THE CONSTRUCTION AND VALIDATION OF AN ARABIC PLACEMENT TEST TO FIRST YEAR STUDENTS AT THE UNIVERSITY OF MALAYA

by

HASSAN DAHAN

Thesis submitted to The University of Edinburgh for the Degree of Doctor of Philosophy (Ph.D)

1999

Declaration

I hereby declare that this thesis has been composed by myself, and that it does not represent the work of any other person and that every work which has been consulted is duly acknowledged.

....

(HASSAN BASRI AWANG MAT DAHAN)
THE UNIVERSITY OF EDINBURGH
OCTOBER 1999

بسم الله الرحمن الرحيم

I dedicate this thesis to all Arabic teachers

Acknowledgements

I would like to express my sincere appreciation to the following:

1. The University of Malaya's administration for its funding of my studies at the University of Edinburgh;
2. My supervisor, Professor M. Y. Suleiman, from the Department of Islamic and Middle Eastern Studies, for his brilliant supervision, intellectual guidance and resolute support from the beginning until the last day of my study;
3. My colleague, Claire Thomson, for her help with my English and for devoting her time to reading the chapters and making the necessary corrections;
4. My family for their inspiration and support of this study and for their patience during their time away from their home country;
5. The Dean of the Language Centre and the Head of Department of the Faculty of *Uṣūluddīn* at the University of Jordan, the officers at the Ministry of Education and the teachers at schools in Malaysia for granting me permission to conduct a pilot test on their students;
6. My colleagues at the Faculty of Language and Linguistics, Faculty of Education, Faculty of Economics and Administration at the University of Malaya, and Malaysian students in Jordan for their assistance in running the test at the Academy of Islamic Studies and collecting the data for this research;
7. The Director of the Academy of Islamic studies at the University of Malaya for permitting me to use new students at the Academy as a sample for this study;
8. Last but not least those who have in any way helped me to complete this study.

To all of them, many thanks, and may Allah, the Mighty, reward them.

THE CONSTRUCTION AND VALIDATION OF AN ARABIC PLACEMENT TEST TO FIRST YEAR STUDENTS AT THE UNIVERSITY OF MALAYA

Abstract

The issue which has always been discussed by scholars in the area of language teaching and testing is whether the test is valid, i.e. whether it tests what it is supposed to test and whether the test is reliable, i.e. consistent in assessing the candidates. This research attempts to construct and to validate an Arabic placement test for new students at the Academy of Islamic Studies at the University of Malaya in Malaysia. The design of the test was based on the syllabus at the Academy and the test specification, prepared during this research. Four sub-tests are constructed: Reading, Grammar, Writing and Dictation. To ensure the validity of the test, two analyses, internal and external, are conducted. The internal validity analysis is concerned with face and content validity. Three groups of students from different levels of academic background and countries participated in the pilot study for the purpose of internal validity analysis. Modifications were made to some items of the sub-tests at the end of the pilot study. The external validity analysis is concerned with concurrent and predictive validity. The correlation coefficients (r) between the total mark of the sub-tests and one of the two measures for concurrent validity indicate that the relationships are moderate: between .40 and .60. As for predictive validity, the r between the sub-tests and the total mark of the final examination are between .60 and .64: a moderate relationship too. In the analysis of the reliability of the tests using the internal consistency method, the reliability coefficients (r_{xx}) for the sub-tests are very high: ranging between .87 and .90. The correlation analysis between the total score of the sub-tests also indicates a very high relationship: five correlation coefficients (r) are between .70 and .75 and only one correlation has the r of .69. The conclusion of the study states that all four sub-tests prove to be successful in assessing the students' proficiency in Arabic and therefore could be used for the purpose of grouping students for the teaching and learning of Arabic at the Academy.

CONTENTS

TABLE OF TRANSLITERATIONS

(a) Consonants

ء	’
ب	b
ت	t
ث	th
ج	j
ح	ḥ
خ	kh
د	d
ذ	dh
ر	r
ز	z
س	s
ش	sh
ص	ṣ
ض	ḍ
ط	ṭ
ظ	ẓ
ع	‘
غ	gh
ف	f
ق	q
ك	k
ل	l
م	m
ن	n
ه	h
و	w
ي	y

(b) Short vowels

fatḥa	=	a
kasra	=	i
ḍamma	=	u

(c) Long vowels

fatḥa and *alif*	=	ā
kasra and *yā*	=	ī
ḍamma and *waw*	=	ū

(d) Dipthongs

fatḥa and *ya*	=	ay
fatḥa and *waw*	=	aw

(e) Other combinations of sounds

wa-al	=	*wa'l*
fi-al	=	*fi'l*
dhū-al	=	*dhu'l*
ة	=	a, except in construct state (iḍāfa) when it is *at*

List of Abbreviations

AIID	Alpha if item deleted
AIS	Academy of Islamic Studies
ANA	American National Standards Institute
ALS	Arabic language syllabus
APA	American Psychological Association
APT	Arabic Placement Test
CITC	Corrected item-total correlation
CLA	Communicative language ability
CLT	Communicative language testing
CT	Cloze test
CTN.ORG	content and organisation
DE	Distractor efficiency
DI	Discrimination index
EAP	English for academic purposes
EFL	English as a foreign language
ELBA	English Language Battery
ESL	English as a second language
EPTB	English Proficiency Test Battery
ESLPE	English as a second language placement examination
FISL	Federal Islamic School of Labu
GRE	General Record Examination
H_0	Null hypothesis
H_1	Alternative hypothesis
HCE	Higher Certificate of Education
ID	Item discrimination
IEPT	Illinois English Placement
IF	Item facility
IMES	Islamic and Middle Eastern Studies
k	Total number of marks, items, scripts, etc.
K-R20	Kuder-Richardson formula 20
K-R21	Kuder-Richardson formula 21
MC	Multiple choice

MCE	Malaysian Certificate of Education
MECHAN	mechanic
N	Total number of candidates, samples, students, etc.
OHP	Overhead projector
p	Probability
Q	Question
ROC	Rank Order Correlation
r	Correlation coefficient
r_{xx}	Reliability coefficient
SAAIC	Sultan Abdul Aziz Islamic College
SD	Standard deviation
SE*r*	Standard Error of the correlation coefficient
SPSS	Statistical Package for Social Science
TMC	Test methods characteristics
TOEFL	Test of English as a Foreign Language
TGRAM1	Total mark of Part One of the Grammar Test
TGRAM2	Total mark of Part Two of the Grammar Test
TREAD1	Total mark of Part One of the Reading Test
TREAD2	Total mark of Part Two of the Reading Test
TREAD3	Total mark of Part Three of the Reading Test
TTLDIC	Total mark for the Dictation Test
TTLESSAY	Total mark for the Writing Test
TTLFINAL	Total mark for the final test
TTLGRAM	Total mark for the Grammar Test
TTLREAD	Total mark for the Reading Test
UCLA	University of California at Los Angeles
VAR	Variable
VOCAB	Vocabulary

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The statement of the problem

Language testing plays a very important role in language teaching. It is a topic of concern to those involved in education whether they are teachers or those involved in research and administration. Bachman and Palmer, (1996) for example, believe that language tests can be a valuable tool for giving information regarding language teaching. "They can provide evidence of the result of learning and instruction, and hence feedback on the effectiveness of the teaching programme itself" (p.8). Bachman and Palmer (op. cit:8) add that tests

> "...can also provide information that is relevant to making decisions about individuals, such as determining what specific kinds of learning materials and activities should be provided to students, based on a diagnosis of their strengths, weaknesses, deciding whether individual students or an entire class are ready to move on to another unit of instruction, and assigning grades on the basis of students' achievement."

It is unlikely to find any teaching without testing at the end of it. As Heaton (1979:1) stresses "Both testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other." Testing is as old as education itself. As long as there have been teachers they have wanted to know how much their students have learned. Testing therefore provides goals for language teaching as well as an outcome to it. Davies (1990) believes that language testing is considered to be central to language teaching. Davies dismisses some opinions which state that testing is marginal in language teaching. Harrison

(1983) agrees with Davies by denying views saying that testing is quite separate from teaching and learning either in theory or in practice. Valette (1977) mentions three major roles of testing; defining course objectives, stimulating student progress, and evaluating class achievement. Gronlund (1982) notes that testing plays a projecting role in all types of learning programmes. He points out that the main purpose of testing is to improve learning whether the test is of the placement, achievement or proficiency type. For the achievement test for example, Gronlund highlights six basic principles to improve learning (Gronlund, op. cit:8-13): (a) tests should measure specifically what has been set in the syllabus; (b) tests should, because of time restrictions and other constrains, measure a representative sample of the learning tasks; (c) tests should include the types of test items that are most appropriate for measuring the desired learning outcomes; (d) tests should fit the particular uses that will be made of the result, which means that whether the test will be used for the purpose of identifying learning difficulties among students (diagnostic test), placing students in certain level of study (placement test) or for the general achievement at the end of learning (achievement test)...etc., the sample of material included in the test and the difficulty of the test items must be prepared properly to fit the particular uses that will be made of the result; (e) tests should be as reliable as possible and should then be interpreted with caution; and (f) tests should improve student learning.

In view of the importance of testing and the role it plays in promoting learning, the question arises whether a given test has followed the above generally agreed principles. In other words, do questions prepared by teachers and do given marks guarantee a certain standard in the area of testing? Gronlund (1982) believes that despite the widespread use of testing in preparing, guiding and evaluating student

learning, many teachers and test developers are not familiar with the generally established ways of constructing tests. This may be explained on the one hand by the fact that many of them receive little or no training at all in how to prepare or to construct tests. Hence they are not aware of the various criteria, including validity and reliability, in constructing tests. On the other hand, it may be because test developers ignore the importance of test virtues including validity and reliability in constructing tests. Alderson, *et al.* (1996) reveal in their research that of the twelve UK tests reviewed, nine were criticised for failing to provide sufficient evidence of reliability and validity.

Background of the study

Every year about 400 students whose first language is not Arabic enter the Academy of Islamic Studies (AIS) of the University of Malaya where the medium of instruction for Arabic and Islamic subjects is Arabic. The requirements for entering this Academy as stated in Rules and Regulations II in the Amendment of 1980 stipulate the following:

> "Candidates must fulfil the requirements of the Academy which are as follows: candidates must pass with credit the Malay Language paper and pass the Arabic Language paper..."

All of the newly enrolled students have to sit the Arabic placement test at the beginning of the session. The purpose of this test is to group new students for learning Arabic in the first year. This test is administered by the Department of Arabic, Faculty of Languages and Linguistics at the University. Based on the results

of the test, a student will be placed in a particular group for Arabic language learning purposes. My personal experience of the test over a period of years as an Arabic teacher has led me to the conclusion that the test is not dependable because of the following reasons:

(i). The test does not have content validity because it does not test students on those parts of the language which ought to be tested. Heaton (1979:154) stresses that "...the test should be so constructed as to contain a representative sample of the course, the relationship between the test items and the course objectives always being apparent". Looking at the syllabus for the First Year Arabic language at the AIS, we assume that various topics and skills are intended to be covered in the test. However, comparing this syllabus with the test, we can establish that many areas are not covered and therefore the test seems to be missing out some very important elements.

(ii). It is difficult to assess whether the test is reliable or not. From my personal experience of teaching over a period of years in this Academy, no research was undertaken to investigate the reliability of the test. The placement test developers did not conduct any research to ensure the reliability of the constructed test such as pretesting, post-hoc analyses and instruments validation. Davies (1977:57) points out that the reliability of the test will ensure the consistency of results.

> "...An inconsistent test would give meaningless, random results. Before looking at the meaning of results it is important to ensure that they are reliable. Unreliable results can have no meaning apart from their own randomness."

Bachman and Palmer (1990:95) agree with Davies regarding the importance of reliability in the test items. They state: "...unless we can demonstrate that the

inferences we make on the basis of language tests are valid, we have no justification for using test scores for making decisions about individuals."

(iii). My personal teaching experience at the Academy further convinced me that some of the new entrants feel that they were not placed at the right level for the purpose of learning Arabic. Some of them feel, based on what they are learning, that they should be placed at a higher level while others feel that they should be placed at a lower level. The above observations have convinced me that the tests at the AIS cannot be relied on to place new students in groups according to their abilities for learning Arabic.

The purpose of the study

The purpose of this study is therefore to:

I. Construct an Arabic placement test for first-year-in-university students at the Academy of Islamic Studies (AIS) based on the syllabus and test specifications, to be developed in this research; and

II. Validate the items which have been constructed using statistical tools to ensure the test's validity and reliability.

Research hypotheses

Several null hypotheses (H_0) are postulated pertaining to the study:

1. H_0 **1** (for descriptive analysis of the final version of the test): There is no difference between the results of samples in the pilot study and the samples from AIS in terms of central tendency and dispersion;
2. $\boldsymbol{H_0}$ **2** (for correlation analysis): There is no relationship among the performance of the candidates on their scores in the sub-tests;

3. H_0 **3** (for concurrent validity): There is no correlation between the students' self-assessment and the items of the placement test;
4. H_0 **4** (for concurrent validity): There is no correlation between either the students' results for the Higher Certificate of Education (HCE) or the pre-AIS examination and their results for the placement;
5. H_0 **5** (for predictive validity): There is no correlation coefficient between the total score of the placement test (*predictor*) and the total score of the final examination for semester one at the AIS (*sample*).

Statistical elements used in the research

Since the study involves the use of statistical elements, I will briefly describe below these instruments. It should be noted however, that the aim of this research is to present these issues to Arabic teachers whose main area of expertise is not statistics. Therefore the explanation will be by means of words and diagrams rather than through figures, formulae and equations.

1. Central tendency (used in the descriptive analysis): The central tendency refers to the description of the typical behaviour of a group of testees in a test and it shows how the scores of a group cluster together (Brown, 1988, 1996: Alderson *et. al.* 1996). Four statistics instruments are used in estimating central tendency: the mean, the mode, the median, and the midpoint. The mean can be defined as the average scores of the testees. It is obtained by adding up all the individual scores and dividing the total by the total number of testees. The mode refers to the scores that occur frequently. In other words, it shows the score obtained by the majority

of testees. The median can be described as the point of middle value. Its function is to cut the distribution of marks into two. The midpoint refers to the point halfway between the highest and the lowest scores. It is obtained by adding up the lowest and the highest scores of a test and then dividing the total by two. For example, if we have a set of scores, say 1, 2, 2, 4, 6, we may calculate, by referring to the above definitions, that the mean is 3, the mode is 2; the median is also 2; and the midpoint is 3.5.

2. Dispersion (used in the descriptive analysis): Basically, the dispersion is the opposite of the central tendency, i.e. how the testees' scores are spread out around the central tendency. Three elements are commonly used in describing the dispersion: the variance, the standard deviation, and the range. The variance can be defined as the average of the squared differences of testees' scores from the mean (Rowntree, 1991; Brown, 1996). The standard deviation is the square root of the variance. The range is the number of points between the highest score and the lowest score. Using the same example of the distribution of marks above (1, 2, 2, 4, 6), we can calculate the elements of the dispersion as follows: The variance is obtained by firstly, calculating the deviation of every mark from the mean (3), i.e. (-2, -1, -1, 1, 3), and secondly, squaring the deviation of every mark, i.e. (4, 1, 1, 1, 9) and lastly, calculating the average of the total squared deviation, i.e. 3.2. The standard deviation for the above marks is derived by squaring root the variance, i.e. 1.79. As for the range, where the highest score is 6 and the lowest is 1, then the range is 5.
3. Correlation coefficient: A correlation coefficient (symbolised with r) is a statistic which expresses the degree of relationship between two sets of test scores or

variables (Harris, 1988:142). The relationship, which is indicated by a number, can reach as high as 1.0 (for positive correlation) and -1.0 (for negative correlation). Rowntree (1991:160) divides correlation into three types, namely the *positive*, the *negative*, and the *zero* correlations. Positive correlation is when "...the changes in one variable are accompanied by changes in the other variable and in the *same* direction; that is, the larger values on one variable tend to go with larger values on the other". Negative correlation refers to "...the changes between the two variables or values in *opposite* directions. Larger values on one will tend to go with smaller values on the other" (Rowntree, op. cit:160). This indicates that if students score high on one test, they score low on the other, or vice versa. Zero correlation, $r = 0$, refers to "...no clear tendency for the values on one variable to move in a particular direction (up or down) with changes in the other variable" (Rowntree op. cit:160).

4. Item facility (IF) (used in a statistical analysis for objective tests): IF is also defined as item difficulty or item easiness. It is used in examining the percentage of samples who correctly answer a given item using the value from 0.00 (no one answers the question correctly) to 1.00 (all answer the question correctly). This value can be interpreted as the percentage of correct answers for a given item by moving the decimal point two places to the right (Brown, 1996). For example, if the index of IF is .16, this means that only 16% of the samples answered the question correctly. On the other hand, an item with an IF index of .95 would indicate that 95% of the sample responded to the question accurately. It can generally be said that an item with a low IF index means the question is difficult and an item with a high IF index indicates that the question is easy.

5. Item discrimination (ID) (used in statistical analysis for objective tests): The ID refers to the degree to which an item separates high scores from low scores in a test. To do the ID analysis, the samples' result has to be lined up, beginning with their individual item responses and ending with their total scores in descending order. Then two groups will be identified: an upper group and a lower group. "The upper and lower groups are sometimes defined as the upper and lower third, or 33%" (Brown, 1996: 67). Brown (op. cit.) even finds some instances where 25% was used in calculating an ID index. There are many ways of calculating an ID index but one of the easiest ways, as suggested by Alderson et.al (1996) and Brown (op. cit.), is by subtracting the IF for the lower group from the IF for the upper group on each item. For example, if we have IF indices of upper and lower groups of one item, say .80 and .45 respectively, we may say that the ID index for the item is .35.
6. Distractor efficiency (DE) (used in statistical analysis for objective tests): The distractor which refers to an option that is counted as incorrect, is particularly used in multiple-choice items. The primary goal of DE analysis is to examine the degree to which the distractors are attracting students. To obtain the distractor efficiency, the percentages of each option functioning in the question are calculated.
7. Inter-rater reliability (used in determining the consistency of the marking of subjective tests): It refers to the degree of similarity between different examiners in giving marks to the same scripts or oral performances (Alderson.*et al.* 1996). According to Carroll and Hall (1985:121) "...a simple way to check on inter-marker [inter-rater] reliability is to use the ranking method". This method is implemented by giving a number of scripts to two examiners and then asking them

to mark them independently according to the marking scheme. Then, using the following formula, the Rank Order Correlation (ROC) is calculated (the formula was taken from Carroll and Hall, op. cit:119):

$$R = 1 - \frac{6 \times \text{Total } d^2}{n(n^2 - 1)}$$

where: R = Rank-order correlation
Total d^2 = the total of differences squared between two examiners
n = the number of scripts marked

The ROC is determined by the value between 0.00 and 1.00. Therefore, a low value of ROC, say .10, indicates that the examiners are not consistent in giving their marks to the script. In contrast, a high value, say .89, means that they are consistent in giving their marks and they therefore can be considered to have high reliability in marking.

The reference style used in this research

It is my responsibility to clarify here that the reference style used in this research adopts the one established by the American Psychological Association (APA). For example, the American National Standards Institute (ANA) (1977), a member of the APA, states that for *'In-Text Reference'*, different methods are commonly used. One of these methods is "...(2) Use of author-date combinations, keyed to the authors' surnames and to the dates of the publications (for example, ...text (Jones 1974)...)" (p. 32).

The APA (1996:97), in its current publication manual, states that:

> "For a direct quotation in the text, give the author, year, and page number in parenthesis (paragraph may be used in place of page number for electronic text). Include a complete reference in the reference list. Depending on where the quotation falls within a sentence or the text, punctuation differs. When paraphrasing or referring to an idea contained in another work, authors are not required to provide a page number. Nevertheless, authors are encouraged to do so, especially when it would help an interested reader locate the relevant passage in a long or complex text".

The significance of the study

This study highlights the importance of validity and reliability analyses in the construction of a test. Therefore the issue that a test is not valid and not reliable could be overcome in the future.

Limitations of the study

The test which will be constructed and validated is based on the Arabic language syllabus for first year students at the Academy of Islamic Studies of the University of Malaya. The sample population involved in the study was limited to first-year-in-university students, who enrolled during the session of 1998/99. This research will not evaluate other aspects of Arabic that the students may have acquired.

Overview of the chapters

In Chapter One, I present a brief overview of trends in language testing. The topics discussed in this chapter cover issues belonging to the pre-scientific period in

testing, followed by the psychometric-structuralist era, and lastly the integrative sociolinguistic trend. Several examples of test are briefly discussed, thus giving the researcher a general guide as to the way test items should be constructed. I end the chapter with a discussion of three types of test: placement test, proficiency test, and achievement test.

Chapter Two is concerned with the analysis of the tests at the Academy of Islamic Studies (AIS). The main purpose of the analysis is to prove the earlier claim that tests carried out at the Academy lack the important characteristics of a good test. Two types of test, placement tests and achievement tests, are used in the analysis. To prepare the ground for carrying out the above task, I describe the validity and reliability of such test which are the prime considerations in language testing.

In Chapter Three, the focus is on the first part of the research, i.e. the construction of the placement test at the AIS. At the beginning of this chapter, a detailed test specification is laid down to guide the researcher towards the construction of the test. Four elements are discussed: item writing, the content and description of the test, the analysis of behavioural objectives, and methods of scoring. One set of prepared tests is given at the end of the chapter.

Chapter Four focuses on piloting the draft test and the analysis of internal validity. Three sample groups, participated in this study, representing different levels of scholastic, academic backgrounds, nationalities, and proficiency in Arabic. Two types of validity, face validity and content validity, are analysed. Several statistical tools such as item facility, item discrimination, and distractor efficiency are employed for the purpose of internal analysis. As a result of the analysis, some test items have been modified or discarded.

The discussion in Chapter Five begins with the administration of the final version of the test on the real samples. This is followed by three types of analysis: the correlation analysis between the sub-tests, the test's reliability analysis, the external validity analysis of the test which involves concurrent validity and predictive validity.

In Chapter Six, I set the passing marks, which is the last task in the construction of any test. Several procedures are mentioned and the most appropriate methods are selected to assign the grades. Finally the conclusions pertaining to the topic of this research are discussed. In the concluding chapter, the problems of this research together with suggestions for future study are put forward.

1. CHAPTER ONE: REVIEW OF THE LITERATURE

1.1 Introduction

This chapter attempts to discuss the advances made in the area of language testing. The discussion will focus on two major aspects of language testing: trends and approaches in language testing and types of language testing. Since psychometric procedures are involved in interpreting the results of measurement activities, this chapter will also touch on these. Two examples of test of the most recent trend in language testing, dictation and cloze tests, will be discussed. At the end of the chapter, three major types of language test, namely placement, proficiency, and achievement, will be described. The purpose of discussing them is to provide some preliminary information for the construction of a sample of a test in the next chapter.

1.2 Trends in language testing

It is interesting to note here that practices in language testing seem to develop from trends in language learning and teaching and developments in language theory (Upshur, 1972; Davies, 1968, 1977). Upshur, for example, suggests that, "Trends in second-language testing tend to follow trends in second-language teaching, and ...--at least in recent times--trends in second-language testing have tended to follow trends in linguistics" (1972: 435). Language testing has thus become one of the most fruitful areas in which linguistics may be applied in language teaching. Moreover, it has become one of the areas where the relevance of linguistic theory can quickly be tested

in practice, and the difficulties in practice can easily be referred to theory (Spolsky, 1978).

In the light of developments in linguistic theory and language teaching and testing research, Spolsky (1978) divides language testing into three major trends or periods, namely the pre-scientific, the psychometric-structuralist, and the integrative-sociolinguistic. He adds that the trends follow in order but overlap in time and approach. The third trend seems to him to pick up many elements of the first, and the second and third co-exist and compete. Moller (1981) however, adds a fourth trend, namely the communicative, which in Spolsky's view is considered to be the second part of the third trend. "The second part of the trend [the integrative-sociolinguistic] is concerned with the need to test communicative competence" (Spolsky, op. cit: ix). The discussion below will attempt to illustrate these three trends.

1.2.1 The pre-scientific

The pre-scientific trend, which belongs to the traditional approach of language testing, has dominated the language testing field for many years. Harris (1970) gives some examples to show the domination of this trend of test:

> "...the grammar-translation method of instruction provided the basis for most tests of foreign language proficiency developed up to, and during, World War II. Thus, for example, of the 19 tests of modern foreign languages reviewed in *Third Mental Measurement Yearbook*, published in 1949 (Buros, 1949), only three tested auditory perception and/or comprehension, and none attempted to measure oral production. Of the 16 tests of grammar, vocabulary, and reading reviewed in the third *Yearbook*, apparently about three-fourths made use of English ... required translation from English into the foreign language or vice versa..."(p.37)

1.2.1.1 Forms of these tests.

Moller (1982) and Al-Ghamdi (1986) note that forms of test in this period are as follows:

(i) translation from the native language of the learner to the second language, and from the second language to his native language. The texts are normally long, but single sentences are also given in the tests;

(ii) sentence completion;

(iii) dictation;

(iv) oral interview: "Common ingredients are reading a passage aloud in L2, and talking on everyday subjects. Marks awarded are frequently not incorporated into the general result of a language examination but reported separately" (Moller op. cit: 3);

(v) compositions in the target language at an advanced level, usually on general subjects, but frequently on literary topics; and

(vi) selected items of grammatical, textual, or cultural interest.

1.2.1.2 The characteristics of tests in the pre-scientific trend

The characteristics of the tests in this period could be explained as follows:

(i) there is a general lack of concern for statistical matters or for such notions as objectivity and reliability (Spolsky, 1978; Pachinburan, 1985);

(ii) the decision of worth and relevance of the answer of the candidates is usually vested with language teachers. In Spolsky's opinion, "...it assumes that one can and must rely completely on the judgment of an experienced teacher, who can tell

after a few minutes' conversation, or after reading a student's essay, what mark to give" (p. v);

(iii) the assessment in all these types of tests is highly subjective; and

(iv) the proficiency being tested is not clearly identified (Moller, 1982).

In their review of language tests, Vaughan James and Rouve (1973) were unable to find any statement of objectives relating to the examinations set for the General Certificate of Education, Advanced Level, in foreign languages in Britain. They attributed this to: "The refusal to make a sharp distinction between practical and cultural aims in modern language teaching" (p. 59)

In a survey of British-based examinations in English for overseas students, Howell (1975) was also unable to identify clear objectives or specific content in most of the syllabuses and concluded that it is only reasonable that it should be stated what such examinations aim to measure.

1.2.2 The psychometric-structuralist

If the first trend was a characteristic of the decades prior to World War II, the second trend appears during and after it. Harris (1970) states that: "It was during the War years of the early 1940's that the beginnings were laid for a dramatic change in the methods of teaching and, subsequently, of testing foreign languages"(p. 37). In the years during and following the war, rapid scientific and technological advances, the increase in immigration to North America and Australasia together with more and more effective means of transportation and telecommunication are some of the factors which led to greater interchange across national and cultural boundaries and to the need for more people to speak more languages with a degree of proficiency[1]. In the

domain of language testing, the influence of these rapid developments - even in technology - has become inevitable. Spolsky (1978) refers to this trend as 'psychometric-structuralist'.

This trend is marked by interaction and conflict between two groups of experts. The first group of experts are psychologists and testers at the same time. They are responsible for the development of modern theories and techniques of educational measurement for providing 'objective' measures using various statistical techniques to ensure reliability and various kinds of validity. Their first thrust was to show the unreliability of the tests of the pre-scientific trend (Spolsky op. cit.). For example, studies by Pilliner and by others on the marking of essays in 1952 showed how unreliable subjective scores can be. The testers' paramount objective in this period was the statistical measurement of reliability and validity. "Firstly, the shape of all tests, whether predictive or non-predictive, language or non-language, is primarily determined by the need to test the tests for reliability and validity (Ingram, 1968, p. 74)". According to Spolsky (op. cit.), this emphasis had two effects. First, tests like this will require a response in writing and this will limit tests of reading and listening. Second, the items chosen did not reflect newer ideas about language teaching and learning. Test constructors added "scientific" elements to language testing, but left a great number of deficiencies. Carroll (1953) quotes the criticism of language testing that Robert Lado made in the summary of his doctoral thesis:

> "A number of conclusions are reached. They are (1) that a great lag exists in measurement in English as a foreign language, (2) that the lag is connected with unscientific views of language, (3) that the science of language should be used in defining what to teach...." (Carroll, 1953) in (Spolsky, 1978).

Carroll's remark on Lado's dissertation marks most clearly the emergence of the second stage of the 'scientific' trend, which Spolsky (1978) sees as the addition of linguistic principles to language testing.

The second group is constituted by linguistics studies which add notions from the science of language to those from the science of educational measurement. This group, led by Lado, accepts completely the psychometric principles laid down by psychologists as a basis for testing. Lado, for example, explains these principles clearly enough for language teachers and even linguists to understand. However, "...he leaves no doubt that linguists, with their understanding of the nature of language, must be the ones to set the specifications for language tests" (Spolsky, 1978). In view of the marriage of the two fields, i.e. psychometrics and linguistics, the discussion below attempts to look, briefly, at some of the work done by experts in psychometrics and linguistics.

1.2.2.1 Psycholinguistic basis

There are various types of school for the assessment of language command. They range from translating sentences or passages, writing essays, dictation, etc. to choosing the correct answer from a set of multiple choice questions, with each question testing some highly specific point of syntax, morphology, or vocabulary.

Ingram (1978) attempts to deal with some of the interpretations of the term 'psycholinguistics' with a view to relating them to testing practices. This needs to be done, according to her, because the psycholinguistic bases for the practices in testing are not always clear.

According to Ingram (op. cit.), there are two views towards the term psycholinguistics. The first is linked to generative linguistics, and owes its origin to Chomsky (1965, 1968). The second originates from Carroll, who is believed to have been the first to use the term in print in 1953. Carroll sees psycholinguistics as simply a word used to cover any area of joint interest to psychologists and linguists, regardless of theoretical orientation and degree of formality (Ingram, 1978).

Chomsky views language as a infinite set of well-formed sentences. In his view, it is the job of the linguist to describe the universals of language. Further, he believes that the ultimate aim of linguistics is to contribute to the study of the human mind. A grammar, for example, according to him, must not only be descriptively adequate, it must also have explanatory adequacy: it must explain the processes that underlie the functioning of the 'native speaker-hearer' (in Ingram, 1978). This obviously gives psycholinguistics a very central place. Ingram elaborates the Chomskyan interpretation as follows:

> "psycholinguists within the generative framework have accepted Chomsky's presuppositions about the nature of language and language use... They also accept one further characterisation of native speakers--that they possess a language faculty which consists of a competence component and a performance component. (p. 2)

1.2.2.1.1 Psychometrics in language testing

The term 'psychometrics', in the language testing context, generally refers to "any and all utilizations of numerical data and related logical operations in the service of developing, using, and interpreting the result of measurement activities..." (Clark, 1978, p. 15). Clark (op. cit.) in his article, *Psychometric Considerations in Language*

Testing, shows how the psychometric basis is involved in the assessment of language testing. Three types of testing are used in his study to identify aspects of psychometric practice most suited to the development, use, and interpretation of test instruments. These are prognostic, achievement, and proficiency testings. For prognostic measurement, psychometrics uses test instruments to determine, through correlational techniques, the level of language accomplishment that students would be expected to attain if they were to follow particular learning programs (op. cit). In the area of the achievement testing, "a major psychometric concern is that of appropriately sampling [a] content within the confines of an administratively-feasible test" (op. cit: 29). In the psychometricans' point of view, the use of multiple choice techniques is not considered well-suited to diagnostic testing, which is one type of achievement testing, because of statistical and logical factors. Instead completion exercises such as 'fill-in' and other 'constructed responses' techniques are considered more appropriate, even though they lack the scoring speed and convenience of the multiple choice format (Lado, 1961, Clark 1972, 1978, Valette, 1977). In proficiency testing, the psychometric basis involves measuring the student's ability to utilize the tested language. In direct proficiency testing, for example, the psychometric practice will dictate how the test represents real life language-use. It will also ensure that scoring procedures for direct proficiency tests must show actual communicative criteria and a high level of both intra and inter-rater reliability. The use of psychometrics will apply to indirect proficiency tests as well. These tests such as the cloze tests (Taylor,1953), the reduced redundancy test devised by Spolsky and others "...derive validity as proficiency measures through a correlational relationship with

direct proficiency tests, rather than through the face/content validity..." (Clark, 1978: 29).

1.2.2.2 The linguistic basis

In view of the role of linguists in language testing, Lado's contribution has been acknowledged by scholars in this field as that of a pioneer. Cognizant of the value of validity and reliability and the role of statistical techniques to achieve quality in tests, Lado nevertheless insists that linguistic and not statistical techniques should determine the content of a test. According to Lado, the role of statistical techniques would be to serve in "...the refinement of tests, not in the selection of language problems" (Lado, 1957: 5)

His approach towards language testing consists of the separation of the complexities of language into segments. He views language as 'a system of habits of communication'.

> "These habits permit the communicant to give his conscious attention to the over-all meaning he is conveying or perceiving. These habits involve matters of form, meaning, and distribution at several levels of structure..." (Lado, 1961: 22)

As a result of this approach, he applies contrastive analysis in designing tests. A structural contrastive analysis between the learner's mother tongue and the target language serves as a filter as to what is to be tested. According to Lado, structure should not be restricted to grammatical structure (syntax), as seen by test developers in the pre-scientific trend; rather, it covers the different components or levels of the language as a whole (Al-Ghamdi, 1986). This means that testing will cover elements

of both language and linguistic skills. Under elements, Lado includes "...sounds, intonation, stress, morphemes, words, and arrangements of words having meanings that are linguistic and cultural" (Lado, 1961: 25). Each of these elements of language constitutes a variable, summed up under pronunciation, grammatical structure, the lexicon, and cultural meanings respectively. Under linguistic skills, Lado suggests that a teacher deals with the integrated skills of speaking, listening, reading, and writing. These skills can be tested as separate universes, although they never occur separately in language (op. cit).

In an attempt to support and strengthen Lado's view, Moulton (1961: 86) quotes five fundamental linguistic principles "in the forms which came to be the slogans of the day". These five principles are:

(1) a language is a set of habits;

(2) language is speech, not writing;

(3) a language is what its native speakers say, not what someone thinks they ought to say;

(4) languages are different;

(5) teach the language, not about the language.

Harris (1970) stresses that these five principles are still regarded as valid and basic by most applied linguists today and he suggests that these principles may therefore "provide a useful set of criteria against which to judge the extent of linguistic orientation of today's foreign language tests" (p. 38).

The structural linguists, led by Lado, were highly influential, both in terms of getting across their views on the nature of language and in terms of linguistic description. Ingram (1978) writes:

> "The view that spoken language is primary led to a considerable increase in emphasis on spoken skill, which in turn led to the construction of tests for spoken language. Three-quarter of Lado's (1961) pioneering work on language testing is devoted to the description of testing formats which deal with spoken language in some way" (p. 7).

However, some of the ideas put forward by linguists have been exposed to criticism. For example, the choice of the contrastive analysis hypothesis as one of the central assumptions of their work in language testing has made them liable to criticism directed not only at their general theory (for example, Hamp, 1968; Di Pietro, 1971) but also at its application to testing (for example, Upshur, 1962) (Spolsky, 1978). In view of its application to testing, Upshur (1962) criticises the attempt by linguists to make contrastive linguistics universal. Upshur argues that linguists' views of the universalisation of contrastive linguistics has given rise to the implication "...that different language backgrounds, because of their differing linguistic habit structures, will present different transfer problems in the learning of the target language" (p. 126). Upshur believes that contrastive analysis is valid for subjects having the same native language background and the same amount of knowledge of the target language. However, the dilemma occurs when the students have different language backgrounds and have different orders or amount of target language learning. Upshur concludes his argument by saying:

> "It is an impractical method for determining test content when students from many language backgrounds are to be tested, and it is a theoretically invalid method for determining test content when students of a single native language background are to be tested; but the theoretical invalidity of the hypothesis does not mean that it is

> useless in the construction of language tests. It only means that results of tests constructed by this method will be subject to errors of interpretation when viewed as strict measures of comparative learning required of the examinees". (p. 127)

Despite this criticism, the psychometric-structuralist trend has had a significant effect in language testing. One effect of the new psychometric trend was to cause many parties to seek ways of improving reliability and validity in constructing language testing. Multiple marking, analytic marking, structured interviews and guided writing tasks were developed. The major achievement of this trend has probably been the production of a number of well-designed, standardized tests. "Most of the standardized language tests of proficiency have been and are still being constructed in this manner, i.e. the complexities of language are analysed into levels and skills and are tested independently..." (Al-Ghamdi, 1986: 63). Some examples of these tests, in English language, are the MLA Foreign Language Tests for Teachers and Advanced Students, the Test Of English as a Foreign Language (TOEFL) developed by the Educational Testing Service, Princeton, and College Entrance Examination Board Achievement Tests in various languages. Spolsky (1978) is of the view that these tests are all good-quality tests in this tradition, widely and confidently used to measure student progress and programme success. The TOEFL, for instance, is now given four times a year at 112 centres in the US and 260 overseas (op. cit.). Other tests are the Edinburgh Proficiency Test Battery (EPTB) devised by Alan Davies (1964, 1965), 'Michigan Test' devised by Upshur and John (1961) and the English Language Battery (ELBA). constructed by Ingram and used by the University of Edinburgh.

1.2.2.3 Characteristics of psychometric-structuralist tests

Tests in this trend can be characterised by:

(a) stress on discrete linguistic points, i.e. phonology, grammar, lexicon, and integrated skills, listening comprehension, and writing;

(b) bias towards testing receptive skills and testing linguistic elements through receptive skill tasks;

(c) extensive use of objectively scored tests; emphasis on greater test reliability and validity; and

(d) control by linguists and psychometricans in construction of tests.

1.2.3 The integrative sociolinguistic approach

Language testing has seen the development of another trend: the integrative sociolinguistic approach. The word *integrative* was first used by Carroll (1961) who raises the issues of the ineffectiveness of discrete-points tests which has been discussed in the section above on the psychometric-structuralist trend. The word *sociolinguistic* is associated with trends in contemporary linguistics which stress the importance of a sociolinguistic approach to the construction of language assessment procedures. Spolsky (1978) attributes the word integrative to the language competence trend, which is connected to various views in psycholinguistics. He adds that this trend "...is based on a belief in such a thing as overall language proficiency, and a feeling that knowledge of a language is more than just the sum of a set of discrete parts". (p. viii) The sociolinguistic trend is ascribed by Spolsky to the 'communicative competence trend', "...it accepts the belief in integrative testing, but insists on the need to add a strong functional dimension to language testing". (op. cit:

viii). However, Valette (1977) and Davies (1978) attribute integrative tests to 'global tests'.

As stated above, the issue of the ineffectiveness of discrete-points tests was first mentioned by Carroll (1961). He argues that discrete structure tests fail to meet a number of basic criteria for measurement of language knowledge. He stresses therefore, the need for tests which do not focus on structural and lexical items only, but also on the overall communicative ability of the testees. It is the combination of these two features, the structural and the communicative, which gives rise to the term integrative and where one pays attention not to specific structural or lexical items only, as in the structuralists' approach, but also to the 'total communicative effect of an utterance'. The discussion below will attempt to look briefly into the reasons behind the appearance of this trend.

1.2.3.1 Integrative approach and sociolinguistic foundation.

It is interesting to note that when Caroll and others were criticising works by structuralists and behaviorists in designing tests, the teaching and learning of language at that time were in some ways dominated by these two groups. Scholars from structuralist groups like Fries (1945) and Lado (1961) claim that learning language is a matter of mastering the sound system, the form, and the structural devices of the language. For example, Fries states that someone has learned a language when :

> "... he has thus first, within a limited vocabulary, mastered the sound system (that is, when he can understand the stream of speech and achieve an understandable production of it) and has, second, made the structural devices (that is, the basic arrangements of utterances) matters of automatic habit." (p.3) in Spolsky (1973: 164)

It is obvious from this statement that Fries arrives at this position after first showing the inadequacy of the notion that knowing a language means knowing its vocabulary (Spolsky, 1973). However, he maintains a related notion, that knowing a language involves knowing a set of items.

Fries's view of language learning has been challenged by some scholars. Spolsky (1973) argues that if we consider the learning of language to be a matter of listing items and listing patterns for arrangement, then we can say that someone has learned the language when he manages to list down the 'sound system' and the 'structural devices'. He adds that "...the list of phonemes would be quite small, no more than sixty or so items, so that it would be quite easy to test each item..."(p. 165). However, according to Spolsky, the criteria for knowing a language are usually determined quite differently. Spolsky adds that statements such as: "I know enough French to read a newspaper," or "He can't speak enough English to ask the time of day" refer to language use and not to grammar or phonology. This implies that in investigating someone's ability in language, we will not usually say : "He has memorised all phonetic elements of the language" or "He has memorised all grammar topics". All of this suggests the impossibility of characterizing levels of knowing a language in linguistic terms, that is, as mastery of a criterion percentage of items in a grammar and lexicon.

In analysing Fries's description of learning a language, Spolsky (op. cit.) concludes that there are many reasons why Fries's approach, which sees that learning language is a matter of mastering the sound system and the structural devices, has not proved successful. "...one of the fundamental reasons is that it [Fries's approach] fails

to take into account two vital truths about language, the fact that language is redundant,[2] and the fact that it is creative" (p. 167).

In order to decide whether a learner knows enough of the language, some approaches have been put forward. One approach is to give him a language-using task to perform.

> "A more promising approach would be to work for a functional definition of levels: we should aim not to test how much of a language someone knows, but to test his ability to operate in a specified sociolinguistic situation with specified ease or effect" (Spolsky, 1968, p. 93).

Another approach is what Carroll (1961) calls the 'integrative approach test': to attempt to characterise in linguistic terms the knowledge of the language required to function in the linguistic knowledge which correlates with the functional ability (op. cit). Spolsky (1967) applies one method based on two assumptions:

> (i) "that there is such a factor as overall proficiency in a second language, and
> (ii) that it may be measured by testing a subject's ability to send and receive messages under varying conditions of distortion of the conducting medium." (p. 39)

Based on these assumptions, Spolsky devised a test in 1967 consisting of fifty sentences which are controlled for vocabulary and sentence structure, recorded on tape and to which white noise is added with varying signal to noise ratios. The testees have to write down what they heard (Moller, 1982). The test shows that the more noise was added, the more mistakes were made; it also shows that some non-native

speakers did as well or better than native speakers when there was no added noise. "This is to be explained by the non-native's inability to function with reduced redundancy..." (Spolsky, 1973, p. 170). However, with the assumption that the sentences in this test represent a sample of the language, there are certain practical difficulties. Firstly, as Spolsky (1968) notes, it is not clear whether errors are ascribed only to the noise level. Secondly, since all the sentences in the test are not related to each other, the context of each has to be decided by the testee. Hence the sentence becomes difficult to recognise, and the whole sentence may be misunderstood. Thirdly, the scoring procedure for this kind of test is not an easy one (Moller, 1982).

Explorations have been made to refine this type of test. At Indiana University, where the test was developed, testees were presented with alternative sentences on the answer sheet and had to select the one they thought they had just heard (Moller, 1982). A further refinement has been developed at Bar-Ilan University in Israel. Testees listen to a tape recorder with white noise added and at intervals the tape recorder stops. The testees then read four alternatives and choose the one they think they have just heard (Whiteson, 1972, and Seliger, 1975).

The original test, which was devised by Spolsky in 1967, was revised again in 1977. New items were written in such a way as to increase the face validity of the test in terms of the situation in which it is most often used (Gaise, Gradman, and Spolsky, 1977). While the items in the first test were discrete sentences, the items in this test appeared to be more contextual:

> "Items which presumably take place in the classroom deal with both general

classroom procedures and specific subject matter taken from a variety of fields - linguistics, history, science, music, and so on. Other items deal with situations which could take place in stores, the bank, the post office, and offices" (op. cit., p. 53).

In order to determine the effect of added noise, one form of the test was prepared without noise. Moreover, the subjects were a mix of native speakers of English, non-native speakers no longer enrolled in remedial English, and non-native speakers still enrolled in remedial English. Gaise, et al (1977) note that the findings of the revised test show: (1) the use of background noise seems to have had little effect on the measurement of overall proficiency for the testees. "While there was a slight improvement on performance on the ten sentences without noise..., the improvement appeared to be minimal" (op. cit: 55); (2) the test does differentiate effectively between native and non-native speakers of English; and it differentiates between non-native speakers as well; (3) the most significant finding of the study, according to Gaise *et al*, may be that the best way to give a dictation test is to contextualise it.

The following discussion considers two types of tests which are very popular in the integrative-sociolinguistic trend. They are the dictation and cloze tests.

1.2.3.2 Dictation

Dictation can be defined as:

"...a technique used in both language teaching and language testing in which a passage is read aloud to students, with pauses during which they must try to write down what they have heard as accurately as possible" (Richard *et al.* 1985:81)

Another definition of dictation is given by Taylor (1980). He views dictation as:

> "(i) reading a passage aloud, (ii) dividing the passage into phrases suitable for committal to STM (short-term memory) and re-reading phrase by phrase with gaps long enough for subjects to record the preceding phrase in writing, (iii) optionally re-reading each phrase as being written, and (iv) re-reading the whole passage in (i)" (p. 88).

Dictation is one of the oldest techniques known for testing progress in the learning of a foreign language (Stansfield, 1985). It is believed that until the end of the Middle Ages dictation was used to transmit course content from teacher to pupil in the first language classroom. Dictation was also the common way of writing a book in the medieval *scriptorium*, a room in a monastery where a master usually dictated to a group of writers (Kelly, 1969).

It is important to note here that the use of dictation as a foreign language testing device cannot be separated from the history of the trends in foreign language teaching and learning. Because of this, dictation may become popular with a particular trend and may disappear or become unpopular with other trends. "It is like a mini-skirt in fashion; once it was liked by many people, then it disappeared, and recently it has become popular again" (Fachrurrazy, 1989: 48).

When the grammar translation method was used widely in teaching foreign languages, dictation was used extensively. At that time, dictation was used as a technique for teaching and testing a foreign language, along with translation essays, oral interviews, sentence completion, and questions on appreciation of literature and culture (op. cit. 1989). It is assumed that writing accurately from dictation would have to be taught to students just as it was taught to writers in the Middle Ages (Stansfield, 1985). The use of dictation as a foreign language testing device was

almost totally rejected during the reign of the natural method. This method, which became popular in the second half of the nineteenth century, discouraged the teaching of reading and writing in the foreign language. Gouin (1894), in (Stansfield, op. cit: 121), who was one of the pioneers of this method, indicates clearly his rejection of the use of dictation:

> "No more dictation lessons. This deplorable exercise is severely interdicted... It would be better simply to copy; the pupil at least would not make mistakes, and to copy he does not need a master. During the time that he scribbles and blots on a page under dictation, he might assimilate it and read it over twenty times. Therefore we have no more corrections, no more recitation, no more dictation" (pp. 331-32).

Dictation regained popularity when the direct method was in favour at the very end of the nineteenth century. This method was considered by its proponents to be more scientific than the natural method since it included the teaching of phonetics. "Phonetic dictation is very stimulating to pupils, and serves as a useful test of their acoustic powers" (Sweets, 1899) in (Stransfield, op. cit: 122). In one study conducted on French and American pupils and college freshmen, Brown (1915) found how dictation lessons can improve students' learning of language, especially in writing. In his observation of French classes, Brown noticed that the French approach to the teaching of composition consisted of the daily use of dictation from primary school onwards. Because of this, French children of ten or twelve could write a difficult passage with almost perfect accuracy. Brown constructed an English passage and dictated it to twenty-eight pupils of eleven and twelve years old and found that eleven of them wrote the passage without error. He then conducted a dictation test with the

same passage on five hundred boys and girls of the same age in eighteen different schools in the United States. Of the total number of papers obtained only eleven were perfect. With the same passage, Brown conducted a dictation test on five hundred college freshmen and found only forty seven of those papers were perfect. Brown concludes:

> "This comparison and others of a similar kind that I have made are sufficient to convince one beyond doubt that the French boy of eleven or twelve has gained materially over the American boy of the same age in writing language accurately" (p. 61) in (Stransfield, op. cit: 123)

Dictation became popular again in the 1930s and 1940s when the reading method was used in foreign language teaching. However, "during the 1960s, the use of dictation began to decline sharply because of the development and widespread adoption of the audio-lingual method..." (Stransfield, op. cit: 123-24). This method, which was influenced by two schools of thought: the structuralist (linguistics) and the behaviourist (psychology), argues that dictation appears to lack any relation to the type of behaviour that human beings normally use to communicate, appears to measure few aspects of language, and does not test word order and vocabulary since they are given (Lado 1961). In addition, dictation was regarded as generally both uneconomical and imprecise, a very indirect and inadequate test of any important skills and as primarily a test of spelling (Anderson 1953, Somaratne 1957, and Harris 1969). The general view about dictation among textbook writers in Europe during the 1960s was almost the same: they did not give dictation any favourable treatment (Stansfield, 1985). Bennet (1968) and Otter (1968) complain that dictation has little

relation to any real life activity and seems to be an extravagant use of examination time. Because of this, dictation, during this period, was strongly critised - especially from the standpoint of language testing.

In the 1970s, with the development of the integrative socio-linguistics trend, interest in dictation returned again. The integrative approach, which involves the testing of language in context, does not separate language skills into neat divisions; instead, this approach is often designed to assess the learner's ability to use two or more skills simultaneously (Fachrurrazy, 1989).

> "The integrated skills thus involved in the test of dictation include listeningcomprehension, the auditory memory span, spelling, the recognition of sound segments, and a familiarity with the grammatical and lexical patterning of the language" (Heaton, 1979: 185)

Scholars of this period like Oller, Irvine, Atai, Valette, and Cohen conducted several series of studies on dictation and concluded that it was a good measure of listening comprehension and overall language proficiency. While Lado (1961) and others, as cited above, describe dictation as a poor measure of language proficiency, Oller (1971) and others have proved, through scientific studies, that dictation can be used as a device for testing overall language proficiency. Based on a theory in the field of cognitive psychology, it is assumed that dictation can tap the learner's internalized grammar of expectancies at work during the listening process (Oller and Streiff 1975). They also noted that it is easy to construct, administer, and score. Oller (1971) notes that with dictation, the student is tested for his ability in three things: "...[he or she must] (a) discriminate phonological units, (b) make decisions concerning

word boundaries in order to discover sequences of words and phrases that make sense, i.e. that are grammatical and meaningful, and (c) translate this analysis into a graphemic representation" (p. 259). Oller (1971) conducted one study on the English as a Second Language Placement Examination (ESLPE) in the University of California at Los Angeles (UCLA) in 1968. The examination consisted of five parts. These were composition, vocabulary, phonology, dictation and grammar. Oller found that "...dictation correlated more highly with each other part of the test than did any other part" (p. 254). In other words, when the correlations between each section and each other section were rank-ordered, the dictation came out first in every possible category. In another study, Oller intercorrelated scores of about eight hundred new foreign students on the ESLPE between 1969 and 1971, and came up with an average correlation between dictation and total score of .91. In another study on a group of students of English as a foreign language in Iran, Irvin, Atai and Oller (1974) correlated their scores on dictation with the scores on the various sections of the TOEFL and found that dictation correlated best with the student's listening comprehension and total TOEFL score. In another earlier study on dictation, Valette (1964) reported that dictation results can be used as an alternative to the final examination. She notes, "For students possessing minimal experience with dictée, the dictée can validly be substituted for the traditional final examination..." (p. 434). In a later study, Valette (1967) reports that she also noticed a high correlation (.90) between scores on a dictation and combined listening, reading, and writing scores on a German examination. One of the latest studies using dictation as a device for testing foreign language proficiency demonstrates that this method is a reliable and valid language testing technique (Fouly and Cziko 1985). In this study, the researchers

construct what they call a 'graduated' dictation test, which contains fourteen segments ranging in length from 2 to 21 words with the shortest segments at the beginning of the text and the longer ones towards the end. The entire duration of the test is approximately 20 minutes. To obtain validity, the scores of the graduated dictation test were correlated with scores of other tests: Illinois English Placement Test (IEPT) and three sub-tests of TOEFL, i.e. listening comprehension, reading comprehension, and structure and written expression, as well as the total TOEFL scores (op. cit.). Fouly and Cziko comment that:

> "Moderately high and statistically significant ($p < .01$) positive correlations, all within a fairly narrow range of .50 to .60 were found between graduated dictation scores and scores obtained on all other measures.... It is of interest to note that the dictation test yielded these consistently high correlations using only 14 items, far fewer items than contained in any of the other measures (the next lowest number of items being 30 for the cloze test)" (p. 561-62)

In terms of reliability, the internal consistency of the 14 segments showed a moderately high reliability coefficient of .85. (op. cit.).

The survey cited above draws the following conclusions. Firstly, dictation is widely used in both language teaching and testing. Secondly, since its validity is so widely accepted, dictation is highly recommended for use on locally constructed proficiency tests utilized for placement purposes (see Harrison 1983 as an example) and dictation is beginning to appear on standardized tests of language proficiency (see Lombardo 1981 as an example). Thirdly, as cited by Oller and others above, dictation correlates positively with more than one aspect of testing in foreign language

(phonology, vocabulary, grammar). Thus dictation is considered a valid measure of overall proficiency[3].

1.2.3.3 Cloze test

The word 'cloze' is derived from the word "closure" (Taylor, 1953). One of the definitions of cloze tests is:

> "A method of intercepting a message from a transmitter (writer or speaker), mutilating its language patterns by deleting parts, and so administering it to receivers (readers or learners) that their attempts to make the patterns whole again potentially yield a considerable number of cloze units" (op. cit: 416).

The cloze test has been used for various purposes. In a through survey of the literature, Alderson (1978) notes that the use of the cloze test was initiated sometime in 1897 by Ebbinghaus. Alderson reports that Ebbinghaus used a 'gap filling' technique for the measurement of intelligence. Alderson has also reported that researchers such as Brown (1910), Ballard (1920) and Hamilton (1929) studied the use of the cloze test including sentence completion and gap filling. Anderson (1970) however, argues that gap filling and sentence-completion test are not the same as the cloze test. "In both blank-filling and sentence-completion tests, words for deletion are chosen quite subjectively. With cloze procedure words are deleted mechanically" (p. 180).

In the 1950s,Taylor (1953 and 1956) used the cloze test as a device for measuring the readability of texts. Later, Taylor (1959) suggested that this kind of test can also be used as a measure of reading comprehension in the learning of foreign languages.

However, as in the case of dictation, the use of the cloze test began to decline sharply because of the development and widespread adoption of the audio-lingual method in the teaching and learning of English as a Second Language. The development of psycholinguistic testing as a series of discrete structures contributed to the decline of the use of cloze test as a test device. The cloze technique was not even mentioned in many standard textbooks on language testing (eg. Lado 1961; Valette 1967; Harris 1969), nor was it discussed in the most widely used language teaching manuals (eg. Lado 1964; Brooks 1964; Rivers 1968) (Oller and Conrad 1971). Only a few studies in the sixties and early seventies were conducted on the use of the cloze test. The study has used students for whom English is a second or foreign language as a sample (Anderson 1970). However, towards the end of 1960's, the cloze test started to be used as a result of the boost given to it by Darnell (1968) and Anderson (1969) who recognised its potential as a measure of proficiency in testing English as a Second Language. In Darnell's (1968) study using a cloze test and modified scoring system, he reported satisfactory reliability and high correlation (.83) with the TOEFL total test score. Since then research into the cloze test has been pursued extensively and the cloze test became one of the most talked about tests in the 1970s and 1980s (Pachinburavan, 1985). Allen (1968), for example, proposed using fill-in-blank tests instead of the multiple-choice types. Estrada (1969) tested the difficulty of various sentence types for Navajo children using the cloze method. Crawford (1970) used the same method to measure the reading comprehension in English of Spanish-speaking American children, and to determine appropriate levels of instructional materials for them.

Many scholars regard cloze procedures as a good, valid and reliable test. Anderson (1970) values the cloze test as "...one of the most promising techniques to emerge in recent years for measuring comprehension and reading difficulty" (p. 181). In English as a Second Language (ESL) proficiency testing, the use of the cloze technique is applicable and the cloze test can be used in the placement of non-native speakers of English and in the diagnosis of their special language problems (Oller, and Conrad, 1971; Oller 1973; Oller, Atai and Irvine 1974; Aitken 1977; Stubbs and Tucker 1974). Aitken (1977) who has constructed, administered, and scored over a thousand cloze tests to ESL students has found that this kind of test is extremely simple, and valid in the proficiency test. The cloze test is also found to have very high concurrent validity. In Shohamy's study (1983) using texts in Hebrew, she found high concurrent validity between cloze test and oral interviews.

Alderson (1979) stresses that three factors are important in the construction of cloze test to ensure its reliability and validity. These are:

(1) The selection of the cloze text

It is important for test constructors to note that the results from a cloze test based on a carefully selected text would correlate highly with other tests while a randomly selected passage would correlate less highly and discriminate less well (Johnson 1980). Johnson adds that it "...seems likely that, given a highly selected passage, variation of the deletion rate might not affect the results to a statistically significant degree, while this would be less likely if a passage were selected at random" (p. 179). Test constructors need also to consider several factors before choosing a text, for testing purposes, for a particular group. Among these factors are intellectual content, cultural content, linguistic difficulty, register and level of

formality, and idiosyncrasies of style, eg. lists of items and a high proportion of idioms, proper names, and numbers (op. cit.). The stress in this type of test is on content validity. If these factors are not taken into consideration in choosing a text, the cloze test might not satisfy the basic prerequisites for a claim of objectivity and hence would be qualitatively subjective (House 1977). Another factor which may affect the objectivity of cloze test is bias. Bias arises when:

(i) the selection is consciously or unconsciously influenced by human choice;

(ii) the sampling frame which serves as the basis for selection does not cover the population adequately, completely or accurately (Moser and Kalton, [1971], in Johnson 1980).

Two types of text are usually used for cloze purposes: narrative texts and expository texts. It is believed that the former are easier to score than the latter because of the availability of narrative schemes which are an inseparable part of narrative texts, and that knowledge of these schemes will help readers to interpret a narrative text (Bullock and Lantolf 1987).

> "...with expository texts, the reading task may be more difficult because comprehension is constrained by the reader's ability to cull and process information from the microstructure of the text. That is, as far as anyone can determine, there are no schemas for expository texts" (op. cit. p. 97).

To prove their argument, Bullock and Lantolf developed two cloze tests in English based on a narrative and an expository text, and administered them to three groups of students. The result of these experiments showed that the subjects performed better on the narrative than on the expository text.

(2) Deletion type and rate

There are two common forms of deletion patterns of cloze tests. The most commonly used is called a fixed-ratio method which consists of deleting every *n*th word of a prose passage (Oller 1972; Irvine, Atai and Oller 1974; Stubbs and Tucker 1974; Pachinburavan 1985). In this pattern, a variety of word deletion frequency is applied. The most common deletion rate is every fifth word (MacGinitie 1961; Bormuth 1963; Ruddell 1964; Alderson 1979). It has been found that a deletion pattern of less than every fourth word, or of more than every tenth word, is either unmanageable or impractical to construct (MacGinitie, op. cit.). Studies undertaken of the effect of deletion rates on the mean scores, however, seem to be contradictory. Oller's study (1972) for example, finds that deletions between every fifth and every twelfth word keep results stable. On the other hand, Alderson (1978) claims that deletion rates do change the results. Regarding the number of deletions in any given passage, Pack (1973) suggests that whatever system is chosen, the passage length should be adjusted in such a way as to accommodate about fifty deletions.

Another form is called the variable-ratio method from which the rationale is used as a basis for justifying the deletion. For example, only function words such as prepositions, articles, conjunctions etc. are omitted while in other tests only content words e.g. noun, adverbs etc. are left out. Berkoff's study (1976) (cited in Pachinburavan 1985) shows that content words are more difficult to restore than structural or function words. Henzeli (1979) reports that, ranking by degree of difficulty in restoration, adjectives are the most difficult words to be restored, followed by adverbs, nouns, verbs, prepositions, pronouns and articles in that order. However, Waiman (1979) claims that there is no significant difference between the

restoration of content words and structural words. Some criticisms of the variable-ratio method of deletion are that it is more difficult, seems to be less valid, and may be biased in favour of certain grammatical categories (Oller and Conrad 1971).

(3) Scoring methods

The most common methods of scoring cloze tests are: (i) the verbatim method, i.e. exact word replacement, and (ii) any contextually appropriate replacement which covers several kinds of elements such as grammatically appropriate, semantically appropriate or both together (Anderson 1971, Stubbs and Tucker 1974, Aitken 1977). Anderson (1971) notices that the two methods give the same results and recommends the former one because it is easier in terms of marking. Stubbs and Tucker (1974:240) suggest that the verbatim method is as valid as the other method: "We found a significant, positive correlation (r=.97, p<.01) between scoring for exact versus contextually-appropriate responses". Oller (1972) however, argues that the contextually-appropriate responses method is better for use with students of second language learning. Alderson (1979), taking into account the types of responses a student has supplied, describes five scoring procedures for cloze tests. These procedures are (p. 195):

1. The exact word procedure;
2. The semantically acceptable procedure;
3. The same form class procedure;
4. The acceptable form class, same grammatical function procedure; and
5. The grammatically correct procedure

It is obvious that the last three of these procedures evolved in an attempt to measure grammatical sensitivity.

From this discussion, we can say that the cloze test is regarded as a reliable and valid measure of proficiency. It is easy to construct, requires relatively little time to administer, and is more objective in scoring and in its presentation (Pack 1973; Aitken 1977; Stubbs and Tucker 1974). Even though there are weaknesses in cloze tests as reported by Klien-Braley (1983), numerous researches have confirmed its reliability and validity.

1.3 Types of test

There are four basic types of language test[5]: proficiency tests, achievement tests, diagnostic tests, and placement tests (Hughes,1992; Gronlund, 1982; Harrison 1983). However, some scholars like Heaton (1979) add another type of test called the aptitude test and consider placement tests as part of proficiency tests. The area of the aptitude test, however, is relatively new, and no aptitude measures even in the teaching of English as a second language could be said to have passed the experimental stage (Harris,1988). This section however, will discuss the following types of test, namely placement, proficiency, and achievement. The purpose of discussing these types of test is to provide some preliminary information before the researcher can construct and develop a sample of a test in the next chapter. Other types of test are considered irrelevant to this research, and therefore will not discussed in this chapter.

1.3.1 Placement tests

Placement tests are designed to measure students' ability in the target language where the outcome of the tests could be used to place them at a certain level

of a teaching programme most appropriate to their abilities. Placement tests "...provide an invaluable aid for placing each student at the most beneficial position in the instructional sequence" (Gronlund, 1982, p. 3). This kind of test is "...concerned with the students' present standing, and so relates to general ability rather than specific points of learning" (Harrison, p. 4). General ability here means that the test could be based on some or all of the following: information or narration, integrative tests such as cloze or dictation, something written, and something spoken. Thus this kind of test looks forward to the course the student is going to take.

Gronlund (1982) believes that placement tests do not assess the students' ability only but also the planned syllabus. He suggests that before proceeding with the instruction, teachers need to answer two major questions:

1. To what extent do the students possess the skills and abilities that are needed to begin the planned instruction?;
2. To what extent have the students already achieved the intended learning outcomes of the planned instruction?

According to Gronlund, a placement pretest covering the intended learning outcomes of the planned instruction can answer the second question. Gronlund (op. cit) stresses that if the outcome of the test shows that students have already mastered some of the material the teachers plan to include in their instruction, the teachers need to modify teaching plans, upgrade the syllabus, encourage some students to skip from certain units or be exempted and placed at a more advanced level. Gronlund (op. cit) notes also that placement tests are not always necessary if:

1. the teacher knows very well his or her students' achievement after working with them for a long time;

2. a course or unit of instruction does not have a clearly defined set of prerequisite skills; and
3. some areas of instruction are so new to the students that it can be predicted that none of the students have achieved the intended outcomes of the planned instruction.

Teachers do not always need to prepare their own placement tests. They can buy or get them from any commercial supplier as long as they are sure that the test being considered suits their particular programme. However, Hughes (1992) believes that no one placement test will work for every institution, and the initial assumption about any test that is commercially available must be that it will not work well.

1.3.2 Proficiency tests

This type of test is designed to measure people's ability in certain aspects of language without referring to any training or any instruction they have had before in that particular language. It is not usually related to any particular past or previous course because it is concerned with student's current performance in relation to his future needs (Hughes 1992; Harrison 1983; Davies, 1970,1977). Hughes (1992) stresses that on this basis, the content of a proficiency test does not necessarily rely on the content or objectives of language courses which people taking the test may have followed. It is rather, according to him, based on a specification of what candidates should be able to do in that test in order to be considered proficient. The word 'proficient' here means, in the case of some proficiency tests, having sufficient command of the language for a particular purpose. Hence this test looks forward to "...defining a student's language proficiency with reference to a particular task which

he will be required to perform" (Heaton, op. cit: 164). It could be said from the definition above that proficiency tests are in no way related, for future purposes, to any syllabus or teaching programme. Hence the main concern of the test is whether a student has enough command of test requirements, for example, language skills, to follow the programme or to perform his duties.

There are other types of proficiency test which do not have any occupation or course of study in mind (Hughes 1992). The concept of proficiency in tests of this kind is more general. These tests are normally conducted by independent examining bodies and are usually relied on by institutions or employers to make comparisons between candidates. Some examples of this type of test are the Cambridge examinations (First Certificate Examination and Proficiency Examination) and the Oxford English as a Foreign Language (EFL) examinations (Preliminary and Higher). The function of these tests is merely to show whether candidates have achieved a certain standard in language learning with respect to certain specified abilities.

1.3.3 Achievement tests.

Achievement tests are formal tests, concerned with assessing what has been learned of a known syllabus and administered normally at the end of a course of study (Davies 1977; Heaton 1979; Gronlund 1982; Hughes 1992).

> "Achievement testing plays a prominent role in all types of instructional programs. It is the most widely used method of assessing pupil achievement in classroom instruction, and it is an indispensable procedure in individualized and programmed instruction" (Gronlund op. cit: 1)

There are two types of achievement test: *progress* achievement tests and *final* achievement tests (Hughes op. cit.; Heaton op. cit). The progress test is usually designed to measure to what extent the students have mastered the material taught in the classroom. The test is normally prepared by the class teacher and is just as important as an assessment of the teacher's own work as of the student's own learning. As for the final achievement test, it could be prepared by class teachers or examination syndicates. "Thus the typical external school examinations (Ordinary level or Advanced level in England, and Highers in Scotland), the university degree exams and so on are all examples of achievement tests" (Davies, op. cit: 45).

Although the primary interest of the achievement test is measuring learning outcomes, very often some further use is made of the same test in order to make meaningful decisions about the pupils' future (Davies op. cit.). As with teaching for example, the main purpose of testing is to improve learning, and within this larger context there are a number of specific contributions which achievement tests can make (Gronlund op. cit.). For instance,

> "Achievement tests [could be used to] support and reinforce other aspects of the instructional process. They can aid both the teacher and the student in assessing learning readiness...monitoring learning progress...diagnosing learning difficulties...and evaluating learning outcomes..." (op. cit: 1)

It should be clear from the discussion above that the content of achievement tests must be related to the courses with which they are concerned. The problem, however, is the nature of this relationship. It is a matter of disagreement among

language testers whether the content of the achievement test should relate to course objectives or to a detailed content of a course (Hughes op. cit).

According to some testers, the content of an achievement test should be based directly on a detailed course syllabus or on the books and other materials used in the course concerned, giving rise to what is known as the *syllabus-content approach*. Hughes (op. cit.) argues that "the disadvantage is that if the syllabus is badly designed, or the books and other materials are badly chosen, then the results of the test can be very misleading"(p.11). Hughes gives some examples to illustrate how successful performance on the test may not truly indicate successful achievement of course objectives:

> "...a course may have as an *objective* the development of conversational ability, but the course itself and the test may require students only to utter carefully prepared statements about their home town, the weather, or whatever. ...Yet another course is intended to prepare students for university study in English, but the syllabus (and so the course and the test) may not include listening (with note taking) to English delivered in lecture style on topics of the kind that the students will have to deal with at University. In each of these examples - all of them based on actual cases - test results will fail to show what students have achieved in terms of course objective" (p. 11)

The above disadvantages of the syllabus-content approach in measuring students' performance through achievement tests leads Hughes to suggest that an alternative approach would be to base the test content directly on the course objectives. This approach, according to him, has a number of advantages. Firstly, it compels course designers and language testers to be explicit about objectives. Secondly, this approach can show how far students have achieved the course

objectives and in turn puts pressure on those bodies responsible for selecting books and materials to ensure that these are consistent with the course objectives. Hughes concludes, "...to base test content on course objectives is much to be preferred: it will provide more accurate information about individual and group achievement, and it is likely to promote a more beneficial backwash[6] effect on teaching" (p. 11)

1.3.3.1 Basic principles of achievement testing

To ensure that achievement tests contribute to improved learning and instruction, Gronlund (op. cit.) lists six principles of achievement testing which provide a firm basis for constructing and using classroom tests as a positive force in the teaching-learning process:

(1) tests should measure clearly defined learning outcomes that are in harmony with the instructional objectives;

(2) tests should measure a representative sample of the learning tasks included in the instruction;

(3) tests should include the types of test items that are most appropriate for measuring the desired learning outcomes;

(4) tests should fit the particular uses that will be made of the results;

(5) tests should be as reliable as possible and should then be interpreted with caution, and

(6) tests should improve student learning (pp. 8-13)

Language tests can be distinguished from each other in two ways: in their connection to a known syllabus and in their relation to time scale. A placement test

has a known syllabus and concerns the future; a proficiency test is concerned with assessing what has been learned of a known or an unknown syllabus for future purposes; an achievement test has a known syllabus and concerns the past.

1.4 Summary of Chapter One

In this chapter, I have discussed three trends in language testing starting with the pre-scientific, followed by the psychometric-structuralist, and then the integrative-sociolinguistic. It is obvious from this study that trends in language testing tend to follow trends in second-language teaching (Upshur, 1972; Davies, 1970, 1977). Influenced by the old grammar-translation method, the first trend of language testing focuses on translation, dictation, composition, and oral interview. Tests are essentially examiner based and are scored subjectively. Almost no attention is paid to the basic characteristics of good and sound tests such as validity and reliability. The influence of contrastive linguistics was seen in the second trend, the psychometric-structuralist. As a result of the inauguration of the audio-lingual theory and structural linguistics, combined with a rapid advance in modern technology after the World War two, this trend, led by psychometricians and linguists, tends to show the unreliability of the preceding trend and seeks to utilise contrastive analysis in designing tests. The proponents of this trend view language as a system of habits which involves form and meaning at the different levels of surface structure beginning with the smallest unit, the phoneme and ending with the largest unit of structure, the sentence. The testing associated with this trend was labeled the discrete-point test (Carroll 1961). The development of language testing has seen another trend following Carroll's (op. cit.)

view of the ineffectiveness of the discrete-points test. Caroll argues that testing individuals and isolated items, regardless of their function in communication, may not indicate the testee's ability to use the language appropriately in ordinary language communication. Carroll therefore suggests the use of testing which focuses on the total communicative effect of the message rather than discrete sentence components. The development of the sociolinguistic approach in teaching and learning reflects the importance of the sociolinguistic dimension to language assessment in language teaching. Spolsky calls this trend the integrative sociolinguistic approach.

End Notes:

[1] For example, The Department of Defense and The State Department in America needed to prepare Army Language Proficiency Test for 31 different languages between 1948 and 1951 (Moller, 1982).

[2] For more information of the term 'redundancy' see Hockett (1958).

[3] There are several types of dictation (Oller 1979): (a) standard dictation, (b) partial dictation, (c) dictation with competing noise, (d) dicto-comp, (e) elicited imitation, (f) dictogloss, and (g) combined cloze and dictation.

[4] To see the illustration the nature of the revised test, the directions and examples see Gaise, *et al.* (1977).

[5] Davies (1970, 1977) prefers to use the *uses* of test instead of *type* or *kind*. According to him, kinds or types of tests would include such terms as Oral tests, Writing tests, Comprehension tests, and First Language (L1) tests.

[6] Hughes (1992) defines backwash as "The effect of testing on teaching and learning" (p. 1). Hughes adds that backwash can be harmful or beneficial.

2. CHAPTER TWO: BASIC CONSIDERATIONS IN TEST DESIGN AND THE ANALYSIS OF ARABIC LANGUAGE TESTS

2.1 Introduction

The aim of this chapter is to analyse some of the Arabic language tests, namely the placement and the achievement, used at the Academy of Islamic Studies (AIS) and to validate their face and content validity. Since these types of test are normally based on the designed syllabus at the AIS, the discussion of the syllabus used for the teaching of Arabic at the Academy is also included in this chapter. This chapter also aims to assess the theoretical principles underlying the test construction vis-a-vis current theories in linguistics, syllabus design and language teaching. To prepare the ground for carrying out the above task, we will discuss the validity and reliability that are the prime considerations in language testing.

2.2 Validity

This is the first characteristic of good tests in language testing. It is very important to have a valid test since if a test is not valid for the purpose for which it was designed, then the scores it generates will not mean what they are believed to mean.

2.2.1 The definition of validity

Henning (1987:89) defines validity as follows[1]:

"Validity in general refers to the appropriateness of a given test or any of its

> component parts as a measure of what it is purported to measure. A test is said to be valid to the extent that it measures what it is supposed to measure. It follows that the term *valid*, when used to describe a test should usually be accompanied by the preposition *for*. Any test then may be valid for some purposes, but not for others".

This definition and others allow for degrees of validity: tests are more or less valid for their purposes. This boils down to saying that validity is not an all-or-nothing matter. "This important point means that users will have to use their own, or somebody else's, judgement when deciding, on the basis of evidence, on the relative validity of a test" (Alderson, *et al.*, 1996:170). From this definition, we also derive the point that the validity of a test ensures its meaningfulness. A test is meaningful within the terms of what is wanted from the test in question.

2.2.2 Types of validity

On the basis of the above definition, the validity of a test may be said to concern the following: *What* precisely does a test measure and *how* well does it measure it? A number of types of validation are applied to tests, all of which attempt to answer the above questions. The terms used for explaining the types of validity sometimes differ from one tester to another and consequently lead to confusion. Alderson, *et al.* (1996) state that over recent years the increasing interest in different aspects of validity has led to a confusing array of names and definitions. However, most testers, even if they have used different terms, have identified three main types of validity: *rational* or *content* validity, *empirical* or *concurrent* validity, *predictive* validity and *construct* validity (Davies, 1965 & 1977; Thorndike and Hagen, 1986; Brown 1988, Heaton, 1975; Harris, 1988).

Rational or content validation measures the test's content to ensure it contains a representative sample of the relevant language skills. To put it in a different way, the test developer must answer the question: Is the test a representative sample of the content of whatever the test is claiming to test? Empirical or concurrent validation relies on empirical and statistical evidence as to whether the students' marks on the test are similar to their marks on other appropriate measures of their ability, such as their scores on other tests known or believed to be valid and given at the same time, their self-assessment or their teachers' rating of their ability or any other such form of independent assessment given later (Heaton, 1979, Alderson, *et al.* 1996). Construct validity refers to what the test scores really mean. As Alderson *et al.* ask, if the test is supposed to test the students' ability to use reference and cohesion in writing, does it in fact do so? Brown (1988) adds that to understand construct validity, we need to understand another related concept - *psychological construct*. A psychological construct, according to him, is a theoretical level construct that is given to some human attribute or ability that cannot be seen or touched because it belongs to the brain.

However, research into test validity has progressed where it may be no longer useful to differentiate between content and empirical validity, since both methods of validation may include empirical data. Alderson *et al.* (op. cit: 171) write:

> "Content analyses of tests often include systematic studies of test content, with experts being asked, for example, to rate the test content in various ways, some of which can then be evaluated statistically".

In place of this, a suggestion has been made by some scholars like Alderson *et al.* that the terms *internal* and *external* or *criterion* validity could be used, with the distinction being that internal validity relates to studies of the perceived content of the test and its perceived effect, while external validity relates to studies comparing students' test scores with measures of their ability gleaned from outside the test. An example of research regarding validity, below, shows how the present researchers use the term internal and external. In her research, Kattan (1990) clearly states that concurrent and predictive validity belong to the external or criterion-referenced part while content or rational validity belongs to the internal part. Thus the discussion below will divide these types of validity into two parts: internal validity, and external validity.

2.2.2.1 Internal validity

The most common ways of assessing the internal validity of a test are: (a) *face validation*, where non-testers such as students and administrators comment on the value of the test; and (b) *content validation*, where testers or subject experts judge the test (see Heaton 1979; Davies 1977; Harris 1988). Henning (1987: 172) adds a third component, called *response validity*, "...where a growing range of qualitative techniques like self-report or self-observation on the part of test takers are used to understand how they respond to test items and why".

2.2.2.1.1 Face validity

A test is considered to have face validity if the test items look right to other testers, teachers, moderators, and testees. According to Davies (1977), face validity is not a theoretical concept. Face validity refers to surface credibility or public

acceptability and is frequently dismissed by testers as being unscientific and irrelevant (Stevenson 1985). Many researchers like Heaton (1979), Davies (1977), Harris (1988) and Alderson *et al.*(1996) agree that face validity is not validity in the technical sense, "...and can never be permitted to take the place of empirical validation or of the kind of authoritative analysis of content..." (Harris, op. cit:21). However, its importance should not be underestimated, for if the content of a test appears irrelevant, silly, or inappropriate, knowledgeable administrators will hesitate to adopt the test and examinees will lack the proper motivation to take it seriously for their given purposes. On the other hand, if test takers consider a test to be face valid, they are more likely to perform to the best of their ability on that test and to respond appropriately to the test items (Alderson, *et al.* op. cit.). Heaton (op. cit.) argues with regards to face validity in language testing that language tests which have been designed primarily for one country and have content validity and then are adopted by another country may lack face validity in the second country. "A vocabulary or reading comprehension test containing such words as 'typhoon', 'sampan', and 'chopsticks' ...will obviously not be valid in East Africa no matter how valid and useful a test it has proved in Hong Kong" (p. 153).

According to Alderson *et al.*, there has been increased emphasis on face validity since the advent of communicative language testing (CLT). Many advocates of CLT like Morrow (1979,1986), and Carroll (1980,1985) argue that it is important that a communicative language test should look (have face validity) like something one might do 'in the real world' with language. "Insofar as this is not systematically or rigorously defined then it is probably appropriate to label such appeals to 'real life' as belonging to face validity" (Alderson, *et al.* op. cit:172). Clearly, it is important for

test makers to keep face validity in mind in constructing their tests, though sound methods of test construction should never be compromised merely to satisfy public opinion. We agree with Heaton (op. cit.) when he says that although it is no substitute for empirical data, face validity can provide a quick and reasonable guide to testers as well as a balance to too great a concern with statistical analysis.

2.2.2.1.2 Content or rational validity[2]

"Content or rational validity is the *representativeness* or *sampling adequacy* of the content - the substance, the matter, the topics - of a measuring instrument" (Kerlinger 1973:458). To explain what is meant by content or rational validity, we give the following example from Gronlund (1982:127):

> "...we have a list of 500 words that we expect our students to be able to spell correctly at the end of the school year. To test their spelling ability, we might give them a 50-word spelling test. Their performance on these words is important only insofar as it provides evidence of their ability to spell 500 words. Thus, our spelling test would have content validity to the degree to which it provided an adequate sample of 500 words it represented. If we selected only easy words, only difficult words, or only words that represented certain types of common spelling errors, our test would tend to have low content validity. If we selected a balanced sample of words that took these and similar factors into account, our test would tend to have high content validity".

Gronlund makes clear that a test is always a sample of the many questions that could be asked and content validity is a matter of determining whether the sample is representative of the larger domain it is supposed to represent. Gronlund (op. cit: 127) suggests that test developers can build a test that has high content validity by: "(1) identifying the subject-matter topics and the learning outcomes to be measured,

(2) preparing a set of specifications which define the sample of items to be used, (3) constructing a test that closely fits the set of specifications".

Content validation involves gathering the judgments of 'experts' who will judge the degree to which the items on the test actually represent the elements which form the substance of a test. If the experts agree that the items do not represent what the test is claiming to test, the test developer would have to return to the drawing board. If they agree that the test represents what it claims to test, the test would be considered content valid for the purpose of testing. This is perhaps the most important aspect of content validity where test developers have to rely on judgement by experts. (Davies 1977, Alderson, *et al.* 1996). Brown (1988) however, argues that the judgement of experts is accurate only to the extent that biases do not interfere with their judgement. According to Alderson *et al.* (1996), one way of dealing with a bias is by setting such criteria as content statement, test specifications, formal teaching syllabus, a curriculum, or perhaps a domain specification as a guide. Alderson *et al.* add that better procedures for content validation would involve the creation of some data collection instrument where expert judges would then be told how to make and record their judgment. For example, two scales, the Communicative Language Ability (CLA) Scale and the Test Methods Characteristics (TMC) Scale, have been developed with the help of experts who rate the test according to the degree to which it met certain criteria. The CLA facets were rated on a five-point scale and related to the level of ability in the areas of grammatical, textual, sociolinguistic and strategic competence. The TMC facets related to test items and test passages and concerned the testing environment, test rubric, item type and nature of test input. Bachman *et al.* (1988) have used these two rating scales to find a quantifiable way of comparing

the content of the two test batteries. Clapham (1992) has used the TMC scale to evaluate the content of three reading comprehension tests by asking three EAP teachers to rate aspects of the test input of the test items and reading passages.

One way of building-in content validity is to provide judges with a list of skills supposedly being tested by a given set of test items. Then the judges will be asked to indicate against each item the skill it tests (Alderson and Lukmani, 1989). Items on which there is little consensus will be considered to have low content validity.

A further alternative in building-in content validity is at the design stage where a range of teachers are asked to make judgments about the texts used for given types of test and the sorts of tasks students are going to be required to complete. This approach can even be carried out during the development of test specifications and trial test examples, and it shows how early in the test construction process content validation can start.

However, it should be borne in mind that experts do not always agree with each other. This has caused a dilemma for test developers because they need evidence of the validity of their instruments as quickly as possible. Two possible solutions have been proposed. First, test developers may gather other sorts of evidence for validity: external validity, face validity etc.. Second, experts may undergo training, so that disagreement can be minimised (Alderson *et al.* op. cit.).

Content validity is particularly suitable for achievement tests and it is important in both criterion-referenced and norm-referenced tests (Davies 1977, Gronlund 1982). Content validity of an achievement test shows that the test is closely related to the syllabus and that the test measures the subject matter, topics, and learning outcomes covered during the instructional period.

Content validity could also be useful for placement and proficiency tests. Here the content of the test is shown to be acceptable in relation to the expectations of the learner. According to Davies (op. cit.), an assessment must be made of just what the learners whose proficiency is to be tested need to do with the language, what varieties they must employ and in what situations they must use them.

2.2.2.1.3 Response validity

Response validity is obtained by gathering information on how individuals respond to test items. The information normally comes from learners/test takers on their test-taking behaviour and thought. The idea of having response validation arose as a result of research which revealed interesting insights into test performance through learner-centred accounts (see Grotjahn 1986; Cohen 1994).

> "For example, introspection on a cloze task will show whether the student has to answer an item by using the range of reading skills intended by the test designer, or whether all that is needed is some knowledge of the grammatical structure of the phrase in which the item appears" (Alderson, *et al.* op. cit:176).

Studying learners' or testees' reactions and opinions can also be done during a reading comprehension task which may identify weaknesses in test items and may produce cases where students can get an item wrong although they understand the passage, or get it right although they do not understand the passage (Alderson 1990).

The simplest way to gather introspective data is by retrospection, where, after testees have taken a test, they are interviewed about the reasons why they produced the answers they did (Alderson, *et al.* op. cit.). Kattan (1990) has used this method to validate English language test items which she administered to a group of nursing

students. Kattan validates her test items by asking candidates to answer two questions: (a) What kind of strategies did the respondents use in answering questions? And (b) Did the questions tap the skills perceived in the test design? For the reading skill, the respondents were asked how they used the following strategies in answering reading comprehension questions: scanning, skimming, clarification, simplification, cumulative decoding of text meaning and coherence detecting strategies. In listening skills, respondents were asked about the strategies they used to respond to questions for which answers were mentioned directly and questions for which answers were not mentioned directly. For the oral skill, respondents were asked about the strategies they used in planning and organising structures and vocabulary. For the writing skill, respondents were asked to explain how they planned and organised the paragraph, how they took care of grammar, punctuation, and spelling in their writing, and whether they used information from the reading and listening sections in their writing skill.

2.2.2.2 External validity

Another name for this type of validity is *Criterion-Related Validity* (Gronlund 1982; Brown 1988) or *Empirical Validity* (Heaton 1979; Harris 1988). As the name suggests, this type of validity differs from internal validity in that instead of collecting the internal measures of the test, it aims at collecting external measures at the same time as the administration of the experimental test or some time after the experimental test has been given. One obvious feature of external validity is that it usually refers to statistical or empirical measures.

External validity is obtained by comparing the results of the test with the results of some criterion measures such as: (i) an existing test, believed to be valid and given at the same time; or (ii) the teacher's rating or any other such form of independent assessment given at the same time; or (iii) the subsequent performance of the testees on a certain task measured by some valid test; or (iv) the teacher's rating or any other such form of independent assessment given later (Heaton op. cit.; Alderson *et al.*, 1996).

The commonest types of external validity are concurrent and predictive validity which are established by using the correlation coefficient measure[3]. The discussion below deals with concurrent and predictive validity.

2.2.2.2.1 Concurrent validity

The concept of external validity is perhaps most readily understood through a discussion of concurrent validity which is also known as *status* validity. Concurrent validation involves the comparison of the test scores with some other measure for the same candidates taken at roughly the same time as the test to estimate current performance on some criterion (Alderson *et al.* 1996; Gronlund 1982). Alderson *et al.* add that the other measures may be scores from a parallel version of the same test or from some other test; or the testees' self-assessments of their language abilities; or ratings of the candidate on a number of relevant dimensions by teachers, subject experts or other informants. "For instance, we might want to use a test of study skills to estimate what the outcome would be of a careful observation of students in actual study situations" (Gronlund op. cit. p. 128).

In this connection, we may ask why concurrent validity is a necessary procedure in testing. Gronlund (op. cit.) underlines three good reasons for this. First, we may want to check the results of a newly constructed test against some existing test that is known to be valid. Second, we may want to substitute a brief, simple testing procedure for a more complex and time-consuming measure[4]. Third, we may want to determine whether a testing procedure has potential as a predictive instrument. In addition to these three reasons, we may mention how often examination boards need to bring out regular new versions of tests. These new versions need to be validated and the simplest way of validating them is by ensuring the existence of a correlation index between the scores of the new versions and the existing ones.

A problem occurs when there is no test available for the purposes of concurrent validation. In such cases, we can rely on other tests that are known and used in that particular context, even though their reliability and validity are unknown (Alderson *et al.* op. cit.). However, we need to treat the results of any correlation of the experimental test and test of this type very cautiously indeed (op. cit.). Alderson *et al.* stress that we would not expect the two tests not to correlate at all because both test language. Yet we might not expect a high correlation between the two partly because they are presumably testing different aspects of language ability, and partly because of the possible unreliability and uncertain validity of the other test.

Using teachers' ranking to obtain concurrent validity of the test is as useful as comparing test results with other test scores. Since the teachers have taught their students for some considerable time, they should have a good idea of the students' levels of proficiency and may be able to rank them according to their language ability.

To ensure an accurate ranking, it is suggested that at least two teachers rank the same group of students. The skills to be rated should be comparatively easy, such as 'oral fluency', and not the difficult ones such as the receptive skills of reading and listening (Alderson *et al.* op. cit.).

Another method of obtaining concurrent validity is by correlating the students' scores with teachers' ratings of students' performance[5]. As well as comparing test results with teachers' ranking, it is probably useful to compare them with another measure, i.e. students' self-assessment, though it needs to be noted that students may not be as accustomed to and as accurate on rating their language ability as teachers are.

2.2.2.2.2 Predictive validity

If concurrent validity concerns the application of external measures at the same time as the administration of the experimental test, predictive validity differs from it in that the external measures will be applied some time after the test has been given. In other words, predictive validity is concerned with the use of test performance to predict future performance on some other valued measure called *criterion* (Gronlund 1982)[6].

There are various ways to obtain the predictive validity of the test. The simplest way is to give students a test, and then at some appropriate point in the future give them another test of the ability the initial test was intended to predict. Another way is by using a test to screen applicants to any learning institution and then correlate applicants' test scores with their grades made at the end of the semester or term[7].

Another way of obtaining predictive validity is by gathering opinions from subject teachers and tutors who can rate their students' ability in such skills as their writing ability, their oral communicative ability, and so on. The problem in this regard is that many tutors can only assess their students at the end of session, by which time the students will have had ample opportunity to improve their language (Criper and Davies 1988; Wall, Clapham and Alderson 1994).

An example of a predictive validation study might be the validation of a test of language competence for teacher training students who have to pass the language test before they are allowed to enter teaching practice. Predictive validation of the test involves following up those students, and getting their fellow teachers and their teacher-observers to rate them for their language ability in the classroom. The predictive validity of the test for these students would be the correlation between the results of the language test and the ratings of their language ability in class[8].

2.3 Reliability

The second characteristic of good tests in language testing is *reliability*, which is sometimes termed *consistency*. Reliability is a necessary characteristic of any good test because, for the test to be valid, it must first be reliable as a measuring instrument. If a test is administered to the same candidates on different occasions, and it produces different results, it could be said to be unreliable. This means that for a test to be reliable, it must be consistent in its results.

According to Heaton (1979), the factors affecting the reliability of a test are:

(1) the extent of the sample of material selected for testing. The larger the sample, i.e. the more tasks the testee has to perform, the greater the probability that the test as a whole is reliable;

(2) the administration of the test. This is a very important factor in deciding reliability, especially in tests of oral production and listening comprehension. For example, if the quality of a recording for an auditory comprehension test is good and is played for a group under good acoustic conditions while other groups hear it under poor acoustic conditions, this will make for unreliability;

(3) test instruction: test developers have to make sure that the various tasks expected from the testee are made clear to all candidates in the rubrics; and

(4) scoring the test: this is another important factor that affects reliability especially for subjective tests, which face the problem of marker reliability.

2.3.1 Methods of measuring reliability

There are various methods of measuring the reliability of a test: (a) *test-retest* method, which involves administering the same test twice to the same group of testees with a time interval in between; (b) *equivalent-forms* or *parallel forms* method, which means administering two equivalent forms of the test in close succession; and (c) *internal-consistency* method, which involves administering the test once and computing the consistency of the responses within the test (Gronlund 1982; Heaton 1979; Henning 1987; Bachman 1990; Alderson *et al.* 1996).

2.3.1.1 Test-retest method

This method requires the administration of the same form of the test to the same group of testees after a time interval. The test-retest method is appropriate for

tests such as cloze and dictation since testers find it relatively difficult to obtain reliability using other methods because of the interdependence of the parts of the test. This method is also useful for situations in which it is necessary to administer a test more than once such as for measuring testees' language ability at several different points in time. "The test-retest method may also be the concern of a language programme evaluator who is interested in relating changes in language ability to teaching and learning activities in the programme" (Bachman 1990: 181).

The primary concern with this method is ensuring that testees do not themselves change differentially in any systematic way between test administrations (op. cit.) Two sources of inconsistency, *differential practice effects* and *differential changes in ability* might influence the reliability coefficients (Bachman op. cit.; Gronlund 1982). Practice effects may occur when some testees remember some of the items of the first test, and therefore perform better on the second administration of the test. Changes in ability may occur if testees' language ability improves or declines, causing them to perform differently the second time. For this reason, Bachman (op. cit.) notices that there is no single length of time between test administrations that is best for all situations.

> "In each situation, the test developer or user must attempt to determine the extent to which practice and learning are likely to influence test performance, and choose the length of time between test and retest so as to optimize reduction in the effects of both" (p. 182).

In this regard, Gronlund (op. cit.) suggests that it is important to include the time interval in reporting test-retest reliability coefficients as this makes it possible to

determine the extent to which the reliability data are significant for a particular interpretation[9].

2.3.1.2 Equivalent-forms or parallel forms method

With this method, the reliability of a test could be estimated by examining the equivalence of scores obtained from alternate forms of test. The procedure consists of two equivalent forms of a test that are administered to the same group during the same testing session. The issue here is how to ensure the equivalence of both tests in terms of difficulty, the nature of their sampling, length, rubric, etc. Henning (1987: 81) suggests that to demonstrate the equivalence of tests, the tests must

> "(1) show equivalent difficulty as indicated by no significant difference in mean scores when the tests are administered to the same person and their means are compared using the t-test, (2) show equivalent variance when the variances of the scoring distributions of the two tests are compared for the same sample of persons by means of an F-Max test, and (3) show equivalent covariance as indicated by no significant differences in correlation coefficients among equivalent forms or among correlation coefficients of equivalent forms with a concurrent criterion, all administered to the same persons and compared by means of the t-test" (p.81)

Henning (op. cit.) further explains that, in practice, it is very difficult to satisfy the above conditions of means, variances, and correlations. For this reason, Henning suggests that test developers may attempt to *equate* tests rather than establish *equivalence*. According to Henning, equated tests are tests that produce different scores for the same candidate, but the scores of these tests have been equated. For example, a score of X on one test is equivalent to a score of Y on the other[10].

2.3.1.3 The internal-consistency method

This method requires only a single administration of a test. It is concerned with how consistent test takers' performances on the different parts of the test are with each other (Bachman 1990). For example, performance on different items of a multiple-choice test that includes items with different formats - some with blanks to be completed and others with words underlined that may be incorrect - may not be internally consistent.

One approach to examining the internal consistency of a test is the *split-half* method (Heaton 1979; Henning 1987; Gronlund 1982; Bachman 1990; Alderson *et al* 1996). The split-half method is based on the principle that, if an accurate measuring instrument were broken into two parts, the measurements obtained with one part would correspond exactly to those obtained with the other (Heaton op. cit.). Thus, according to this method, a test is divided into two halves, say, the odd items and the even items, and is administered to a group of examinees. The scores of each half are correlated and the extent to which they correlate with each other will govern the reliability.

The test may be split in a variety of ways. If the test is comprised of items, a convenient way of splitting a test might be to divide it into the first and second halves. The problem with this is that one of these halves may be more difficult since most language tests are designed as 'power' tests with the easiest questions at the beginning and the questions becoming progressively more difficult, which means the principle of equal splitting could be not satisfied (Bachman op. cit.). Another procedure widely used is to ascertain the correlation between the scores on the odd-numbered items in one half and all of the even-numbered items in the other half

(Heaton 1979; Henning 1987; Gronlund 1982). However, if the items are graded according to increasing difficulty, the problem will be the same as the above one. A more accurate operation is to divide the items as follows:(Heaton op. cit: 157)

item	1	4	5	8	9	12
against item	2	3	6	7	10	11

For tests without items such as those consisting of a series of compositions, each group of compositions may be scored separately and correlated with the other in order to establish reliability (Henning op. cit.)[11].

In a subjective test, reliability can be assessed by correlating the marks given by two or more judges or raters to the same student and by correlating marks given by the same judge or rater on different occasions. This procedure of obtaining the reliability of the test is called *inter-rater* reliability. There are two steps in the estimation of inter-rater reliability (Henning 1987.). First, the ratings of all judges must be intercorrelated. Second, the correlation coefficient or average coefficient is adjusted by the matrix formula to make the final reliability estimate reflect the number of raters or judges who participated in the rating of the examinees.

It is important to note here that each of these methods of obtaining reliability provides a different type of information (American Psychological Association, [1974], in Gronlund 1982). Therefore, reliability coefficients obtained with the different procedures are not interchangeable. We need to determine what type of reliability evidence we are seeking before choosing the procedure to be used. The table below summarises the types of information provided by each method (Gronlund op. cit: 133):

METHOD	TYPE OF INFORMATION PROVIDED
Test-retest method	The stability of test scores over some given period of time
Equivalent-forms method	The consistency of test scores over different forms of the test (that is, different samples of items)
Internal-consistency method	The consistency of test scores over different parts of the test

2.4 Analysis of the Arabic language syllabus and tests at the Academy of Islamic Studies (AIS)

Introduction:

This part is concerned with two issues. Firstly, it examines the content of the Arabic language syllabus (ALS) at the Academy of Islamic Studies (AIS) in Malaysia. Secondly, it analyses the present Arabic language test at the AIS focusing on the placement and achievement tests. Though this research is mainly related to language testing, the discussion of the syllabus is inevitable since the construction of the test usually depends on the syllabus to which it relates. As Weir (1993) explains, testing should not be divorced from teaching. In other words, testing must be viewed as an integral part of the learning process and should sample the domain specified by the syllabus.

2.4.1 The Arabic language syllabus (ALS) at the Academy of Islamic Studies (AIS).

The learning of Arabic at AIS is divided into two phases: phase one at the Pre-Academy of Islamic Studies centre and phase two at the AIS itself.

2.4.1.1 The Arabic language syllabus at the pre-Academy of Islamic Studies Centre.

Students study Arabic at this Centre for two years, divided into four semesters, before they are accepted at the AIS. In this Centre, in addition to Arabic, they also study other subjects such as Islamic law (*Sharī`a*), Theology (*Uṣūluddīn*), English and Malay. For Arabic, the syllabus is divided into three parts: Arabic language I, Arabic language II, and Arabic Language III. The syllabus consists of the following areas:

(i) Arabic language I. The syllabus consists of three main topics. They are:

1. Syntax (*al-naḥw*) which covers the following topics: the indefinite noun (*al-nakira*), the definite noun (*al-ma`rifa*) which covers proper names (*al-`alam*), demonstrative noun (*ism al-ishārah*), relative pronoun (*ism al-mawṣūl*), and synarthrous (*al-muḥallā bi al*). This part is allocated three hours per week.
2. Morphology (*al-ṣarf*) which covers derivation of verbs (*taṣrīf al-af `āl*), verbs that are bare of any accessory and verbs that have an accessory (*al-mujarrad wa'l-mazīd min al-af`āl*), weak, strong, and hamzated verbs (*al-mu`tall wa'l-ṣaḥīḥ wa'l-mahmūz min al-af`āl*), the intransitive and transitive verbs (*al-lāzim wa'l-muta`addī min al-af`āl*). This part is allocated one hour per week.
3. Rhetoric (*al-balāghah*) which covers comparison (*al-tashbīh*), object and trope (*al-ḥaqīqa wa'l-majāz*), (*al-majāz al-mursal wa`alāqātuhu*), and (*al-majāz al-`aqli wa `alāqātuhu*). This part is allocated one hour per week.

(ii) Arabic language II covers three topics which are:

1. Reading comprehension (*al-muṭāla`ah*) which covers various subject matters such as politics, economics, social studies, and art;

2. Arabic literature covering prose, poetry, and proverbs. This part is allocated two hours per week; and
3. Translation from and into Arabic. Topics are selected from modern books, magazines, and newspapers. This part is allocated one hour per week.

(iii) Arabic language III focuses on three main topics:

1. Composition (*al-maqāl*). Students write, for a period of two hours a week, at least seven topics in a semester. Topics range from writing letters of invitation, application letters to writing a complete essay on such topics as describing the nature of the world. It is normal practice in teaching composition that students are taught to discuss topics orally;
2. Oratory (*al-khiṭābah*) (1 hour per week). Students are taught how to deliver good sermons on such occasions as celebrating the Islamic new year, farewell and wedding parties, worship in Islam, the role of fasting during Ramaḍan in creating good manners, and the importance of knowledge; and
3. Dialogue (*al-ḥiwār*) (1 hour per week). Under this theme, students are trained to develop dialogues on such topics as daily activities, introducing oneself, festivals, etc. The purpose of this component is to enable students to speak Arabic fluently and spontaneously.

For the remaining three semesters, the areas covered by the syllabus are the same as those covered in Arabic language I, II, and III, i.e. syntax, morphology, rhetoric, reading comprehension, and composition but at a higher level. Taking syntax as an example, the relatively challenging topics such as *inna* and its sisters (*inna wa akhawātuhā*) and annexation (*iḍāfa*) are not taught until the fourth semester. The same applies to morphology and rhetoric where students are taught such topics as

the broken plurals (*jumū `al-taksīr*), *i`lāl* and *ibdāl*; *al-muḥsināt al-ma`nawiyyah* and *al-muḥsināt al-lafẓiyyah* in the last semester. It is clear from the above that in the course of two years, students learn around fourteen topics. This is equivalent to the syllabus followed by students majoring in Arabic language at a university.

2.4.1.2 Arabic language syllabus at the Academy of Islamic Studies (AIS).

Students study Arabic at the AIS for six semesters in a minimum period of three years before they are awarded a degree in either *Sharī`a* or *Uṣūluddīn* or *Tarbiya Islāmiyya* [12]. Degrees in Arabic language and literature are not awarded at the AIS. Arabic is taught for the purpose of helping students acquire other subjects in the Faculty of *Sharī`a*, the Faculty of *Uṣūluddīn* or *Tarbiya Islāmiyya* (the Programme in Islamic Education)[13].

The First Year Arabic language syllabus at the AIS covers the following topics: Arabic syntax, Arabic morphology, writing, speaking, and reading.

(1) Arabic syntax covers the following:

a. the origins of Arabic grammar (*nash'at al naḥw*)

b. the sentence and its parts (*al-kalām wa mā yata'allafu minhu*);

c. declension and indeclension (*al- i`rāb wa'l-binā'*);

d. indefinite and definite (*al-nakira wa'l-ma`rifa*);

e. the nominal sentence (*al-jumla al-ismiyya*);

f. subject and predicate (*al-mubtada' wa'l-khabar*);

g. *inna* and its sisters (*inna wa akhawātuhā*);

h. *kāna* and its sisters (*kāna wa akhawātuhā*); and

i. verbs of appropinquation (*af`āl al-muqāraba*)

(2) Arabic morphology covers the following:

a. the importance of morphology (*ahamiyyat `ilm al-ṣarf*);

b. singular, dual, and plural (*al-mufrad wa'l-muthannā wa'l-jam`*);

c. aplastic and derivative nouns (*al-jāmid wa'l-mushtaqq*);

d. aplastic and inflected nouns (*al-jāmid wa'l-mutaṣarrif*);

e. verbs that are bare of any accessory and verbs that have accessory (*al-af`āl al-mujarrada wa'l-mazīda*);

f. the derivation of strong and weak verbs (*taṣrīf al-af`āl al-ṣaḥīḥa wa'l-mu`talla*).

(3) Writing and speaking cover the following:

(a) Writing:

Topics for writing are divided into two sub-sections which cover religious and social affairs, and everyday events. Under religious and social affairs, students cover the following:

I. cooperation and unity;
II. sacrifice;
III. the importance of calling people to Islam (*da`wa*);
IV. the role of young people;
V. natural panorama scenes which cover the countryside and the magnificence of the world.

Under everyday events, students cover the following:

I. picnics;
II. accidents ; and
III. application letters which cover: (a) an application to a university, (b) an application for a job, and (c) a letter of apology.

(b) Speaking:

Topics for speaking are as follows:

I. library;

II. university;

III. cafeteria;

IV. post office and telecommunications;

V. public transport;

VI. airport;

VII. government offices;

VIII. shops and supermarkets; and

IX. public places.

(4) Reading covers topics with the following objectives:

Through reading lessons, students will be able to practice grammar and to translate into Malay. Students will acquire no less than 250 new vocabulary items through reading and training in the use of dictionaries. Topics for reading are taken from the following resources: (a) *`Izzat al-Nāshi'īn* by Muṣṭafā al-Ghalāyīni, *al-Risālāt* by Aḥmad Ḥasan al-Ziyyāt, *Waḥyu al-Qalam* by Ṣādiq al-Rifā`i, *al-Naẓarāt* by al-Manfalūṭi, *`Abqariyyat al-imām* by `Aqqād, *al-Ayyām* by Ṭāha Ḥusayn, and *al-Ṭarīq al-ṭawīl* by Najīb Kaylāni; (b) Arabic newspapers and magazines.

For the remaining five semesters, the topics covered by the syllabus are the same as those covered in the first semester, i.e. Arabic syntax, Arabic morphology, writing, speaking, and reading, with the addition of Rhetoric starting from the third semester.

2.4.2 Analysis of Arabic language tests at the Academy of Islamic Studies (AIS).

In this section, I will attempt to analyse Arabic language tests administered at the pre-AIS Centre and the AIS. The analysis focuses on two types of test, which are normally used in these centres: placement and achievement.

2.4.2.1 The Arabic language tests at the pre-AIS Centre.

2.4.2.1.1 The Arabic placement test (see Appendix A.1.1: 394-411).

Background:

Every year, since 1983, the Arabic language division at the pre-AIS Centre has administered an Arabic placement test to all new students at the Centre. The test is prepared by a group of Arabic teachers at the Centre. Its purpose is to assess the students' ability in the Arabic language and thus place them into particular groups suitable for their ability. The great advantage of this test is that it uses entirely simple paper and pencil techniques with three multiple choice questions. Listening and speaking are not included in this test.

- Description of the test.

The following is a description of the Arabic placement test at the pre-AIS for the 1996/97 academic year:

- Cover page: The rubric gives candidates the information to answer the questions, the total number of questions, a space for the candidate's name and his or her identification card number. No time-limit for completion of

the test is stated on the cover of the test booklet. No sample of how to answer the questions is given.

- Test content: The test consists of one hundred multiple-choice items, each with a choice of three options. The maximum possible marks on this test is one hundred. Below are the summary of the topics and total items of the test.

Topics:	**total questions:**
◊ Arabic syntax (*al-naḥw*)	82
◊ morphology (*al-ṣarf*)	8
◊ Translation from and into Arabic	3
◊ Vocabulary and the meaning of words	7
Total	100

The details of the Arabic syntax covered by the test:

Topics:	**total questions**
♦ declension (*al-i`rāb*)	47
♦ verbal sentences (*al-jumla al-fi`liyyah*)	7
♦ the diptote (*mamnū` min'l-ṣarf*)	6
♦ the noun of *inna* and its sisters (*ism inna*)	5
♦ the nominal sentence (*al-jumla al-ismiyya*)	3
♦ the subject of a nominal sentence (*al-mubtada'*)	3
♦ the adjective (*al-ṣifa*)	2
♦ the noun of *kāna* and its sisters (*ism kāna*)	2
♦ the direct object (*maf`ūl bihi*)	2
♦ prepositions (*ḥarf al-jarr*)	1
♦ the feminine (*mu'annath*)	1
♦ the numeral (*al-`adad*)	1

♦ the vocative (*al-munādā*)	1
♦ the exception (*al-istithnā'*)	1
Total	82

The declension questions covered by the test:

Topics:	**total questions**
* the exception (*istithnā'*)	7
* the direct object (*mafūl bihi*)	6
* the predicate (*khabar*)	6
* the accusative (*manṣūb*)	4
* the vocative (*munādā*)	4
* the nominative (*marfū`*)	3
* the denotative of state (*ḥāl*)	2
* the adjective (*ṣifa*)	2
* specification (*tamyīz*)	2
* the indeclinable (*mabnī*)	2
* the cognate accusative (*maf`ūl muṭlaq*)	2
* the causative object (*maf`ūl li'ajlihi*)	1
* the adverbial object (*maf`ūl fīhi*)	1
* the subject (*mubtada'*)	1
* the subject of a verbal sentence (*fā`il*)	1
* the imperative verb (*fi`l al- amr*)	1
* the noun of *kāna* (*ism kāna*)	1
* the predicate of *kāna* (*khabar kāna*)	1
Total	47

The majority of questions are related to syntax and most of the questions on syntax are related to declension (*i`rāb*). These represent half of the 86 questions. Some questions are very detailed, which may confuse the new students.

2.4.2.1.2 Analysis of the placement test at the pre-AIS Centre

To analyse this test in terms of validity and reliability, we need to refer to the characteristics of a good test and to the syllabus we have discussed above.

(1) Face validity: As has been discussed above, face validity refers to *surface credibility* or *public acceptability*. We may say that face validity is not fully satisfied by the pre-AIS test because there is no balance in the skills tested. Vocabulary and translation testing are minimally represented. This lack of balance in testing the various language skills may lead to lack of motivation in the students, thus depressing their performance. The test may also have an impact on the students' expectations of the course, leading them to think that the central core of language teaching is the syntactic component of the grammar. Moreover, due to the lack of balance in the skills tested, the result of the test may not represent the actual language ability of the students.

(2) Content validity: As was pointed out earlier, content validity deals with the *representativeness* or *sampling adequacy* of the content - the substance, the matter, the topics - of a measuring instrument. In order to build a test that has high content validity, Gronlund (1982) suggests that test developers need to identify the subject-matter topics and the learning outcomes to be measured.

By comparing the content of the placement test as set out above with the syllabus of the pre-AIS Centre, which has been summarised in 2.4.1.1 above, we may suggest the following observations:

(a) The test has low content validity since it does not represent the content of the syllabus at that Centre. Arabic syntax represents 86% of the total questions, morphology represents 8%, translation and comprehension each represent only 3% of the total questions of the test. If we refer to the above syllabus, we find that the skills which ought to be taught to the students, i.e. syntax, morphology, translation, speaking, reading and writing, are divided equally in terms of time and content. In other words, there is no extra time allocated for syntax, less time for translation, and so on. It therefore becomes difficult to evaluate the testees' overall ability when the test content represents a small aspect of the syllabus, i.e., in this case, syntax or grammar topics. Candidates who obtain lower grades in the test could not therefore be automatically classified as weak in overall ability because other skills such as reading, listening and writing are not included in the test questions.

(b) It may be argued that since syntax represents more than 80% of the total questions of the test, the result obtained from this section of the test could be used, as an alternative, to assess the students' ability in Arabic. However, the following argument shows that even syntax questions have low content validity. With reference to the above data, we observe that none of the questions is related to the syllabus of Arabic I at the pre-AIS centre. Most of the questions are related to the syllabus of Arabic II, III, and IV and some are not related to any section of the syllabus at all. For example, the diptote, which has six questions, and the noun of *inna*, which has five questions, are related to Arabic IV. Nominal sentences and the subject of the

nominal sentence, which have three questions, are related to Arabic II. The section on verbal sentences, which contains seven questions, is not related to any part of the syllabus. Declension, which has forty-three questions, is not related to Arabic I. Seven questions for exception and four questions for the vocative are related to Arabic III, while the direct object and predicate, which have six questions in the test are both related to Arabic II. The accusative and nominative are not related to any explicit syllabus item at the Centre. This can be summarised in Table 2-1 and Table 2-2 below:

Table 2-1: Summary of the test content (1996/97)

Topics	total questions:	relationship to the syllabus:
declension	**43**	**see table 2-2 below**
verbal sentences	**7**	**no direct connection**
diptote	**6**	**Arabic IV**
inna and its sisters	**5**	**Arabic IV**
nominal sentences	**3**	**Arabic II**
subject of a nominal sentence	**3**	**Arabic II**

Table 2-2: Summary of the declension topics

Topics (for declension):	total questions:	relationship to syllabus:
exception (*isthisnā'*)	**7**	**Arabic III**
direct object (*maf'ūl bihi*)	**6**	**Arabic II**
predicate (*khabar*)	**6**	**Arabic II**
accusative (*manṣūb*)	**4**	**no direct connection**
vocative (*munādā*)	**4**	**Arabic III**
nominative (*marfū'*)	**3**	**no direct connection**

(Note: The discussion above does not take into consideration topics that are allocated fewer than three questions because they represent a very small percentage in the total marks of the test.)

(c) Another factor which reduces the content validity of this test is the fact that no norms or validity data, or correlation, are provided. Having taught at the Centre for

some time, I suspect that the test constructors are not even aware of these analytical elements of testing.

It is also my personal experience that the opinions of experts are not solicited in assessing the content validity of the test.

2.4.2.2 Arabic language test at the AIS

2.4.2.2.1 Arabic placement test (see Appendices A.1.2: 412 and A.1.3: 418)

Background:

The Arabic placement test at the AIS was first administered in the early nineties. The format of the placement test during the eighties is not clear since the materials are not available. It is understood, however, that learners were grouped at that time either by referring to their Arabic test results at secondary schools or by placement tests at the AIS. During the early nineties increasing numbers of students entering the AIS from various institutions of learning with different abilities of Arabic have raised the issue of Arabic language assessment. To use Arabic test results from their last schools is felt to be inadequate and unreliable because these students come from different institutions which have different types of assessment. As a result, a placement test was conducted to assess the students' proficiency in Arabic and then to group them according to their ability. The following is a description of two Arabic placement tests for the years 1995 and 1996.

- Description of the test

Two papers were analysed for this research, namely Test Paper One and Test Paper Two. The following is a description of these two papers:

- Test Paper One (1995) (see Appendix A.1.2: 412-17): This test was prepared by teachers at the Language Centre who are also responsible for teaching Arabic language at the AIS. The sole purpose of this test is to place students for the learning of Arabic in the first year at the AIS according to their ability. The content of the test can be summarised as follows:
    - Cover page: the Arabic rubric instructs the students to answer the questions and states the total number of questions; there is also a space for the candidate's name and his or her identification number. The time allocated is one hour. There are no examples of how to answer the questions, an omission which may be due to the large variety of question-types contained in the test.
    - Test content: The test consists of four questions. Marks are not allocated to each of these questions. The details of the test questions are as follows:
        ⇒ Question One relates to writing skills. It requires filling in the blanks with various kinds of word classes such as prepositions, pronouns, verbs, and nouns. There are only five questions under this question and the instructions are not clear.
        ⇒ Question Two, which also refers to writing skills, asks testees to re-arrange the words to make complete sentences. The instructions here are not clear either. There are five nominal sentences in this question. There is a minor mistake in question five which may confuse testees.
        ⇒ Question Three asks the testees to correct the grammar mistakes in sentences. There are five nominal sentences in this question. Question 1 asks testees to differentiate between feminine and masculine items; Question 2 and three require the testees to identify the adjectives appropriate to the described word in terms of gender

and declension; and Questions 4 and 5 are also related to the feminine and masculine. Testees are asked to provide the pronouns appropriate to the described subjects.

⇒ Question Four asks the testees to write a short composition based on a cartoon provided along with the question. The length of the composition is set at one hundred to one hundred and twenty words.

- Test Paper Two (1996) (see Appendix A.1.3: 418-24): This test was prepared by the same group of teachers who prepared the above test paper. The content of the test can be summarised as follows:
  - Cover page: The Malay rubric instructs the testees to answer the questions, and states the total number of questions; there is also a space for the candidate's name and his or her identification number. The time allocated is one hour. There are no examples of how to answer the questions, perhaps due to the large variety of question-types contained in the test.
  - Test contents:
    ⇒ Question One is a cloze test with multiple choice answers. It uses a fixed-ratio method which consists of deleting every *n*th word of a prose passage. In this question, the test constructors use the most common deletion rate, i.e. every fifth word. There are ten blanks in the text and the deleted words range from prepositions and pronouns, to verbs and adjectives.

    ⇒ Question Two is about Arabic syntax. It requires candidates to change the active verbs to passive ones or vice versa. There are five verbal sentences in this question: three sentences are in the active forms and the remaining two are in the passive.

⇒ Question Three relates to morphology. Candidates are asked to fill in the blanks with the correct form of a given word. There are five words in this question: three refer to the infinitive nouns and the remaining two refer to passive participle and past verb.

⇒ Question Four tests students' vocabulary. Candidates have to fill in the blanks in five incomplete sentences with words provided in the box.

⇒ Question Five also relates to vocabulary. Candidates are asked to choose one unfamiliar word from the list. There are five items in this question and the unfamiliar words range from foods and clothes, to the human body.

Every question has ten marks, which brings the total to fifty.

2.4.2.2.2 Analysis of the placement tests at the AIS:

(a) Face validity:

- Paper One :

The use of Arabic for the instructions gives the impression that this paper has face validity. In referring to the questions of the test described above, we establish that with the exception of Question Four, most of the questions are very easy for students at the university level.

- Paper Two:

The use of Malay for the instructions is not appropriate. Since Arabic is the medium of instruction at the AIS, Arabic should be used especially in the examination. In terms of content, with the exception of question three which relates to morphology, the questions are at the same level as those in Paper One. This boils down to saying that if a candidate scores high marks in this test, this does not necessarily show that he

or she displays an adequate competence in Arabic and therefore could be exempted from the course.

(b) Content validity:

In terms of content validity, these two tests may be considered valid if they can be shown to test what they are supposed to test. The analysis below determines whether or not the contents of each test are in line with the syllabus described above.

- Paper One:

As was pointed out earlier, items in Questions One and Two relate to writing skills: both questions test students' ability to construct sentences. The issue here is whether gap-filling and re-arranging the words to construct sentences are among the activities for the teaching of writing at the AIS. The syllabus (see 2.4.1.2 (3) (a) does not include these types of activities in the teaching of writing skills. The syllabus focuses almost exclusively on writing essays which are at a higher level than gap-filling and re-arranging the words to construct sentences. For this reason, it is difficult to prove that if candidates obtain higher marks for Questions One and Two, then they are good at writing. This means that these two questions have low content validity.

Question Three has low content validity too. Most of the items in the question are not related to the syllabus. The summary below shows the relationship between the items in the questions and the syllabus:

Items tested:	**Relationship with the syllabus**
feminine and masculine	no syllabus connection
adjective	no syllabus connection
pronouns	no syllabus connection

subject and predicate related to Year One syntax

Question Four, which refers to composition, clearly relates to the syllabus. Even though the topic and consequently the content of the question has no relation whatsoever with the syllabus described above, this question is still considered valid for the testers to assess students' ability in writing.

- Paper Two:

Question One, which is a cloze test, could be said to have content validity. The items in this question are related to the reading skills outlined in the syllabus above. However, a problem still arises if we compare closely the items in the question with the materials used in the syllabus for teaching reading. All references used for teaching reading are classical texts while the text used in question one is a modern and simple text. Question Two, which consists of active and passive verbs, has low content validity because the questions are not among the syntax topics in the syllabus described earlier. The same applies to Question Three, which refers to morphology. The items in the question are infinitive nouns, passive participle and past verb; none of these questions relate to the syllabus. Although Questions Four and Five relate to vocabulary, which is included in the syllabus, it is questionable whether the range of items and the cognitive level of words tested are enough to assess the students' knowledge of vocabulary.

We may say here that the test developers did not identify some of the subject-matter topics and the learning outcomes to be measured when they constructed these test papers.

2.4.2.2.3 The Arabic achievement test (see Appendix A.1.4: 425-438)

Background:

Since the start of the academic year 1980/81, the Arabic language division at the Language Centre has administered an Arabic achievement test to students at the AIS at the end of instruction. The test is prepared by the same teachers who also administer the Arabic placement test described above. The main purpose of the test is to certify competence and assign grades to students. The test is typically broad in scope and attempts to measure a representative sample of all of the learning tasks included in the instruction. The example given below is the 1995/96 achievement test paper.

- Description: The content of the test can be summarised as follows:
    - Cover page: The Arabic rubric instructs the candidates to answer the questions, and the total number of questions. The time allocated is three hours. There are no examples of how to answer the questions due to the large variety of question-types contained in the test. The maximum mark is seventy five only. The remaining twenty five marks are generated from a speaking test, which is administered on a separate occasion. Unfortunately, the items for the speaking test were not available to the researcher.
    - Test content: The test consists of three parts: Part One refers to knowledge of Arabic; Part Two refers to language skills; and Part Three consists of multiple-choice questions.
    - ♦ Part One: Knowledge of Arabic (*al-`ulūm al-`arabiyya*) (see pp. 427-28)

This part consists of two sections: syntax and morphology. Fifteen marks are allocated to this part: ten marks for syntax and the remaining five marks for morphology.

◊ Section One: Syntax.

⇒ Question One asks candidates to provide three sentences: the first sentence should relate to the verbal sentence in which the verb and the agent occur with presumptive vowels (*al-ḥaraka al-muqaddara*); the second sentence should contain an indeclinable accusative of place (*ẓarf al-makān al-mabnī*); the last sentence is the nominal sentence and the predicate must be a defective noun (*al-ism al-manqūṣ*) to which the first person pronoun is annexed.

⇒ Question Two, which contains two sub-questions, covers declension (*i`rāb*): (a) candidates are asked to identify the last vowel of the abbreviated noun (*al-ism al-maqṣūr*) and the weak perfect form (*al-fi`l al-muḍāri` al-mu`tall*); (b) candidates are asked to give examples of declinable verbs (*al-mu`rab min al-af`āl*).

⇒ Question Three asks candidates to identify, from the given Quranic verses, the case classification of the pronouns, i.e. they are nominative (*marfū`*), or accusative (*manṣūb*) or genitive (*majrūr*).

◊ Section Two: morphology.

⇒ Question One refers to augmented verbs (*al-af`āl al-mazīda*). Candidates are required to re-write the five verbs in the question leaving out the augmented letter elements (*al-ḥurūf al-zā'ida*).

⇒ Question Two asks candidates to identify, from a selection of underlined words, types of aplastic nouns (*al-asmā' al-jāmida*). There are five items in this question.

⇒ Question Three, which contains five items, asks candidates to differentiate between aplastic nouns and derived nouns (*al-mushtaqq*).

♦ Part Two: language skills (see pp. 429-432)

This part has two sections: Section One refers to reading comprehension and translation; Section Two refers to writing skills.

◊ Section One has five questions and each question is allocated five marks. With the exception of Question Five, all questions are related to a text in the question paper.

⇒ Question One consists of the reading of a text followed by five comprehension questions.

⇒ Question Two asks students to explain the inflection of five underlined words in the same text. The inflection ranges from direct object, adjective and predicates, to the subject of a nominal sentence.

⇒ Question Three asks candidates to vocalise the third paragraph of the same text.

⇒ Question Four asks candidates to translate the fourth paragraph of the text into Malay.

⇒ Question Five asks candidates to translate a Malay sentence into Arabic.

◊ Section Two: writing skills. Candidates are asked to write an essay of approximately 200 words in length on one of five topics. The topic themes range from descriptive essays such as the description of fasting in Ramaḍan (in letter format), the characteristics of a caller (*dā`i*) to Islam, and a description of a journey, to general topics such as the role of young people

towards the country and life in Malaysian society. Fifteen marks are allocated to this part.

- Part Three: objective questions (see pp. 432-38).

This part, with a total of twenty marks, consists of twenty multiple-choice questions with a choice of five options for each question. The content of the questions ranges from syntax and morphology to language skills such as translation, vocabulary and gap-filling. The list below summarises the details of the test questions:

Topics:	**Total Questions**
(i) syntax:	
declinable and indeclinable nouns (*al-mabnī wa'l-mu`rab*)	7
definite article (*alif lam al-ta`rīf*)	1
definite noun (*al-ma`rifa*)	1
(ii) morphology:	
imperfect verb (*al-fi`l al-muḍāri`*)	1
strong verb (*al-fi`l al-ṣaḥīḥ*)	1
transitive verb (*al-fi`l al-muta`addī*)	2
extra letters (*al-ḥurūf al-zā'ida*)	1
(iii) language skills:	
translation from and into Arabic	2
gap-filling	2
vocabulary	1
re-arranging words to form a sentence	1
Total:	20

2.4.2.2.4 Analysis of the test

(a) Face validity:

On the surface, the test appears to have face validity, an impression strengthened by the use of Arabic in the instructions. In referring to the questions of the test described above, with the exception of some questions, the questions seem to meet the minimum standard of the examination requirements at university level. Thus the result of the test can be said to show the students' actual achievements in Arabic.

(b) Content validity;

To determine the content validity of the above test, we will compare the content of the test with the Arabic syllabus at the AIS discussed earlier (see 2.4.1.2), and we will also compare the content of the test with the principles underlying the development and use of achievement testing.

The above test measures clearly-defined learning outcomes that are in harmony with the instructional objectives. The test measures specific skill types which the students are expected to demonstrate at the end of the learning process. These include: grammar, covering syntax and morphology; reading skills, including comprehension of a set of texts; correct vocalisation and translation; and writing skills, focusing on writing an essay.

In terms of the sampling adequacy of the syllabus content, we could say that the test measures a representative sample of the syllabus. It does not leave out any skills which need to be tested in the syllabus. The list below summarises the relationship between the content of the test and the syllabus mentioned earlier.

Test content :	**Relationship with the syllabus:**
Part One: knowledge of Arabic	
Section 1: syntax:	
Q 1: verbal sentence	related
Q 2: declension	related
Q 3: declension	related
Section 2: morphology:	
Q 1: extra verbs	related
Q 2: nouns that are incapable of growth	related
Q 3: aplastic nouns	related
Part Two: language skills	
Section 1: reading skills:	
Q 1: reading comprehension	related
Q 2: inflection of words	related
Q 3: vocalisation	related
Q 4: translation into Malay	related
Q 5: translation into Arabic	related
Section 2 : writing skills:	
Q 1: writing an essay	related
Part Three: objective questions:	
Arabic syntax	related
Arabic morphology	related
language skills	related

The spread of marks could also be said to be properly allocated: 15 marks are allocated for syntax and morphology; 25 marks to reading skills; and another 15 marks are allocated to writing skills. However, marks for the multiple choice questions are not equally divided. More than 70% of the 45 marks are for grammar and only 30% are for language skills. This, however, does not have very much influence on the candidates' total score; as a result the test can be said to have content validity.

Since the test represents the sampling adequacy of the syllabus content, it can be used for assigning grades or certifying mastery of the instructional objectives. Hence, the results obtained by the candidates accurately show the level of their performance in Arabic. This is another argument in support of the content validity of the above test.

Since the results of the test represent the sampling adequacy of the syllabus, it can also be used to improve student learning. The testers can determine, from the results of the test, areas in which the candidates need help to improve their learning; they can provide feedback on the candidates' test performance as soon as possible after testing; and they can suggest specific aspects of performance which should be improved.

To conclude, we may say that, at this stage, we are able to analyse the above tests only according to their internal validity, i.e., face and content validity. It is difficult, however, to analyse the above tests in relation to concurrent, predictive and construct validity due to the lack of relevant information and data. With regard to concurrent validity for instance, these tests have never been correlated with other tests to check concurrent validity. With regard to predictive validity, these tests are not

geared to predict performance on some other valued measure called criterion. For this reason, no such correlation coefficient can be obtained for the tests, and therefore we cannot assess the highest possible index for perfect positive and negative relationships. The same applies to construct validity which, in fact, is more difficult and complicated to establish than other types of validation.

2.5 Summary of Chapter Two

In this chapter, I have analysed some samples of the tests used at the Academy of Islamic Studies (AIS) to determine the direction of the test construction at the Academy. I have also discussed, in brief, the Arabic syllabus at the AIS. The discussion of this syllabus is necessary since the construction of a test usually depends on a syllabus to which it relates. To prepare the background for the analysis, the discussion of the theoretical concept of validity and reliability, which are the prime consideration in language testing, was put forward. Even though some aspects of validity and reliability were not applied during the analysis of the test items in this chapter, the description of these was essential for two reasons: firstly to inform the reader as to aspects of both validity and reliability that were missing from the current tests, and secondly to prepare the groundwork for the draft test, which will be constructed in the next chapter. The analysis of the test questions for the test papers at the pre-AIS and at the AIS Centre indicated that the construction of the tests, especially for the placement test, still followed the traditional or classical method. The influence of grammar analysis, especially in the test from the pre-AIS Centre, revealed the emphasis placed on grammar by the designers of the syllabus. In terms of

face and content validity, the majority of the test items for the placement test were not related to the designed syllabus at the AIS and this caused the test to have low face and content validity. Having discovered these weaknesses, it is hoped that in the construction of the draft test for the research experiment in the next chapter, the above errors will be avoided.

End Notes:

[1] Other definitions of validity may also be found in Brown (1988), Heaton (1975), Davies (1977), and Harris (1988).

[2] Some testing specialists make no distinction between content and face validity, but consider them to be synonyms (Magnusson, 1967).

[3] To understand how to obtain the external and predictive validity, we need to know what a correlation coefficient is:

> "A correlation coefficient symbolised as (*r*) indicates the degree of relationship between two sets of measures. A positive relationship is indicated when high scores on one measure are accompanied by high scores on the other; low scores on the two measures are associated similarly. A negative relationship is indicated when high scores on one measure are accompanied by low scores on the other measure" (Gronlund 1982, p. 128).

The extreme degrees of relationship that can be obtained between two sets of scores are indicated by the following values:

1.00 = perfect positive relationship
.00 = no relationship
-1.00 = perfect negative relationship

[4] Alderson *et al.* add that "...the other test may not be easily available, or it may be too expensive, or too long for practical use, or it may be a secure test which can be made available only for the purposes of validation but not for regular use by the institution" (p. 178).

[5] The form below might be used by teachers to rate their students' performance: (Alderson *et al.* p. 179)

How would you assess each student on a scale of 1 to 5 for each of the following skills: grammar, writing, speaking, overall language proficiency?

Student	Grammar	Writing	Speaking	Lang. Proficiency
01				
02				
03				
etc.				

The 1 to 5 scale may be a very simple one such as:

1. weak
2. Fairly good
3. Good
4. Very good
5. Like a native speaker

[6] Davies (1977) adds that predictive validity is established by a statistical procedure:

> "Predictive validity ...is established by a statistical procedure... The correlation relates the test scores to an acceptable criterion which is predicted and which is quantifiable. The [experimental] test is the predictor: it shows its predictive validity in its relation to its future criterion which it predicts" (p. 60).

The estimation of this type of validity is usually expressed in terms of correlation coefficients like those commonly used in estimating concurrent validity, with 1.00 and -1.00 as the highest possible index for perfect positive and negative relationships.

[7] "For example, one might administer a university entrance exam to a group of students at the time of their entry into university. One might then proceed to collect grade-point averages (GPAs) for each of these students after each successive year of university study. Finally, one would correlate admissions exam scores with successive annual GPAs to obtain predictive validity of the admissions exam" (Henning 1987, p. 97). However, Alderson *et al.* (1996) argue that the result of any correlations are obscured by the fact that, on the one hand, the class of grade point averages (GPAs) reflects not only language ability, "but also academic abilities, subject knowledge, perseverance, study skills, adaptability to the host culture and context, and many other variables" (p. 181). On the other hand, GPA-type predictive validity will tend to be artificially low because the sample has been truncated on the exam under consideration.

[8] The expectation of correlation coefficient between the experimental test and the external measure is unusually high. For predictive validity for example, it is common for test developers and researchers to be satisfied when they achieve a coefficient as low as +.3 (op. cit.).

[9] A formula for this method might be expressed as follows (Henning 1987):

$$r_{tt} = r_{1,2}$$

where, r_{tt} = the reliability coefficient using this method
$r_{1,2}$ = the correlation of the scores at time one with those at time two for the same test used with the same persons.

[10] The usefulness of this method is as Bachman (op. cit.) observes: "Like the test-retest approach, this is an appropriate means of estimating the reliability of tests for which internal consistency estimates are either inappropriate or not possible" (p. 182-3). Bachman justifies that:

> "It is of particular interest in testing situations where alternate forms of the test may be actually used, either for security reasons, or to minimise the practice effect. In some situations it is not possible to administer the test to all examinees at the same time, and the test user does not wish to take the chance that individuals who take the test first will pass on information about the test to later test takers. In other situations, the test user may wish to measure individuals' language abilities frequently over a period of time, and wants to be sure that any changes in performance are not due to practice effect, and therefore uses alternate forms" (p. 183)

The procedure for calculating reliability using parallel forms is the following (Henning 1987):

$$r_{tt} = r_{A,B}$$

where, r_{tt} = the reliability coefficient
$r_{A,B}$ = the correlation of form A with form B of the test when administered to the same candidates at the same time

Reliability coefficients determined by this method calculate errors within the measurement procedure and consistency over different samples of items. (Gronlund, 1982).

[11] Another formula is by the Spearman-Brown prophecy formula which has been simplified by Gronlund (1982) as follows:

$$\text{Reliability of total test} = \frac{2 \times \text{reliability for 1/2 test}}{1 + \text{reliability for 1/2 test}}$$

For example, if we obtained a correlation coefficient of .70 for two halves of a test, the reliability for the total test would be computed as follows:

$$\text{Reliability of total test} = \frac{2 \times .70}{1 + .70} = \frac{1.40}{1.80} = .77$$

[12] For Islamic Education programme, students will be doing double major when they graduate from AIS, i.e. the teaching of Islamic Education and the teaching of Arabic language.

[13] The programme of Islamic Education is not called *faculty* because it is a joint programme between the AIS, the Faculty of Education, and the Faculty of Language and Linguistics in the University of Malaya.

3. CHAPTER THREE: TEST SPECIFICATION AND TEST CONSTRUCTION

3.1 Introduction

The aim of this chapter is to design and construct a test specification and a draft of the Arabic language placement test which will be used in the research experiment. To prepare the ground for this task, this chapter will discuss some basic steps in preparing the test specification including the general purposes of the test, test outline, the types of test needed, the level and range of item difficulties, and the number of items in the test. Finally, I will construct a draft of the Arabic placement test based on the specifications. Four major aspects of the preparation of the draft test will be discussed, namely item writing, the content and description of the test, the analysis of behavioural objectives and methods of scoring.

3.2 Rationale

The review of literature in Chapter One has revealed that there are various trends in language testing. This literature helps the researcher to determine the most appropriate trend to follow in the experiment. In addition, the analysis of some test samples in the AIS in Chapter Two will be used as a guide to investigate the most appropriate types of test needed.

3.3 Test specification

A test specification is a detailed document, and is often used for internal purposes in the examining body on a confidential basis. A test specification can be

defined as the sum total of the qualities and characteristics that the test should possess (Tinkelman 1971; Alderson *et al.* 1996). According to Bachman and Palmer (1996: 176), two parts can be included in the test specification:

> "(i) the task specifications for each type of task..., and
> (ii) the characteristics that pertain to the structure of the test: the number of items, the salience of parts/tasks, the sequence of parts/tasks, the relative importance of parts/tasks, and the number of tasks per part".

Test specifications constitute a detailed document, often used for internal purposes for the test developers and for those who need to evaluate whether a test has met its aims. Tinkelman (1971: 47) is of the view that the form of the test specifications should be so complete and so explicit "...that two test constructors operating from these specifications independently would produce comparable and interchangeable instruments...".

In order to develop test specifications for the draft paper test at the AIS, I will use some of the basic steps suggested by Tinkelman (op. cit: 47), and Ebel (1979)[1]. These steps are:

(1) defining the general purposes of the test

(2) preparing the test blueprint or outline

(3) planning the types of items to be included in the test

(4) planning the level and range of item difficulties

(5) planning the number of items in the test and its parts

3.3.1 Defining the general purposes of the test

In order to define the general purposes of the test, the following questions will be tackled:

(a) What specific areas of abilities are to be measured?

The test is expected to measure students' proficiency in Arabic which covers language skills, vocabulary acquisition and grammatical accuracy in accordance with year one syllabus in the AIS.

(b) What sort of learner will be taking the test?

The test is intended for the new intake of students who have been accepted to study for the academic year 1998/99 at the AIS.

(c) How are the test scores to be used?

The test scores are to be used to assess students' ability in Arabic and thus place them in particular groups in accordance with their ability. The test scores will also be used to predict students' proficiency in Arabic at the end of the period of instruction.

(d) How long will the test be?

The length of time for the designed test will be not more than three hours in one testing period during the university's orientation week. If it is found that because of a full testing schedule during that week, the administration of the proposed test will be divided into two parts: the first part with one and a half hours of testing time and the second part with one hour testing time.

(e) Will equivalent forms be needed?

Equivalent forms may be indicated if some students miss the initial testing.

The immediate equivalent forms will be the learners' scores of Arabic tests from their last secondary school or the equivalent.

From the answers to the above questions , the statement of general purposes of the draft test may be drawn up as follows:

> Competence in Arabic, as defined for the purposes of this test, consists of knowledge of Arabic grammar, listening skills, reading skills, writing skills and the acquisition of vocabulary for Year One syllabus in the AIS. No attempt is made to include oral skills. This test is to be administered to new students who have been accepted to study at the AIS for the academic year 1998/99. The test scores are to be used to assess the students' ability in Arabic and thus place them in particular groups according to their ability as well as to predict students' proficiency in Arabic at the end of the instruction. The test must be designed not to require more than three hours of administration in any testing period. The equivalent form will be obtained only if any student misses the test.

3.3.2 Preparing the test blueprint or outline

The statement of general purpose seldom supplies test constructors with an adequate basis on which to begin test preparation. The purpose of the test blueprint or outline is therefore to define for the test constructors the scope and emphasis of the test (Tinkelman, op. cit.). The test blueprint can be defined as the plan of stratification that is then followed in drawing up the test sample. There are various reasons why test constructors need the outline of the test. One of the reasons is because a test is a work sample. Thus the role of a carefully prepared outline is to draw a line that all the test contents represent the syllabus. This can be summarised as follows:

(A) The specific purpose of the test

The test is a placement test: its specific purpose is to assess new students' level of Arabic language ability so that they can be placed in the appropriate group or course. Students are placed according to their rank in the test results so that, for example, the students with the top scores go into the top group. The content of the test will be based on the syllabus taught at the AIS as well as unrelated material but at a level equivalent to the syllabus concerned. Since the content of the test covers various skills, students' ability in different skills such as listening and writing will be identified. This means that a student could conceivably be placed in the top reading group, but in the bottom listening group, or some other combination. The test also has the purpose of deciding whether students need to attend a preparatory course or whether they could be exempted from such a course.

(B) The content of the test

The term *test content* has been used to cover both the subject matter of the test and the type of ability that is being tested (Tinkelman op. cit.). This part will establish not only the different topics of subject matter to be covered in the test but also the types of behaviour to be elicited with regard to each area. The details of these two are discussed below:

(i) Topics covered in the test.

For Arabic syntax, Arabic morphology, writing and reading, the syllabus at the pre-AIS and the First Year Arabic syllabus at the AIS are used as the main framework (see 2.3.1. (a) and (b) in Chapter Two for details of the syllabus). For writing and reading skills, the language content is general, referring forward to the course the students are going to take. The language content will not be specified in very much

detail because the candidates are new and come from different learning backgrounds. For the listening test, the materials will be taken from other sources since this skill is not covered in the syllabus.

(ii) Analysis of behavioral objectives

The analysis of behavioral objectives is important because it determines which activities and skills should be appraised in the test. In addition, it should faithfully reflect the objectives of the instruction (Gronlund, 1982). In this connection, a useful tool is the Taxonomy of Educational Objectives by Bloom *et al.* (1956)[2]. There are six taxonomy categories which can be summarised as follows[3]:

(a) knowledge for identifying, naming, defining, describing, listing, matching, selecting, and outlining;

(b) comprehension for classifying, explaining, summarizing, converting, predicting, and distinguishing;

(c) application for demonstrating, computing, solving, modifying, arranging, and operating;

(d) analysis for estimating, separating, ordering, and inferring;

(e) synthesis for combining, formulating, designing, composing, and revising; and

(f) evaluation for judging, critiquing, comparing, justifying, concluding, and discriminating.

When the subject matter of the test and the type of ability that is to be tested have been selected and clearly defined, a two-way chart - which is called a table of specifications - will be prepared. The table relates the outcome of the subject matter and indicates the relative weight to be given to each of the various areas. From this table, the readers can determine the balance of the test contents and the type of the

ability assessed. However, it is important to stress here that this does not mean that there should be a uniform distribution of the taxonomy categories in the table of specifications. Rather, the balance should depend on the subjects and areas of study. With regard to these matters, Tinkelman (1971: 55) says that:

> "Behavioral objectives, such as 'knowledge of definitions' or 'knowledge of generalisations' often are more appropriate for certain content areas than others. In social studies tests, map-reading skill is a common behavioral objective... The test constructor should not hesitate to make adjustments in the blueprint based on logic and reason, while ever mindful of possible confounding effect."

3.3.3 Planning the types of items

The test uses various types of items ranging from multiple choices, dictations, cloze tests, true-false, to writing short essays. The draft distribution of every skill and topic is as follows:

(i) Listening Test: the listening test contains multiple-choice items and dictation. For the Dictation Test, it is a combined skill with the Writing Test

(ii) Reading Test: the Reading Test, which includes a vocabulary test, contains three types of item: cloze test, the true-false test with correction, and multiple-choice items which include multiple-response variation

(iii) Writing Test: the Writing Test consists of dictation and writing a short essay

(iv) Arabic grammar (syntax and morphology): item types for Arabic grammar are true-false with correction and detecting errors in sentences in the multiple-choice format.

3.3.4 Planning the level and range (distribution) of item difficulty

With regard to the level of item difficulty, Richardson (1936) in Ebel (1979: 89) suggests that "...a test composed of items of 50 percent difficulty has a general validity which is higher than tests composed of items of any other degree of difficulty". In the same manner, Gulliksen (1945) in Ebel (op. cit: 90) stresses that "...in order to maximize the reliability and variance of a test the items should have high intercorrelations, all items should be of the same difficulty level, and the level should be as near 50 percent as possible".

Ebel (1979) suggests two ways in which this matter can be approached. The first is to include in the test questions or problems that are answerable by any students. Ebel adds that with this approach, most of the students can be expected to answer the majority of the questions correctly. The questions are very effective in discriminating the various levels of abilities of the students - best, good, average, weak, and poor.

The second approach is to construct tests on the basis of their ability to reveal different levels of proficiency among the students tested. Ebel agrees that this approach requires a preference for somewhat more difficult questions. Taking as an example the open-ended questions: if 32 percent of the examinees answer an item correctly, the item is said to have a difficulty index of 32 percent or .32, usually indicated by the letter *p* which means percentage passing.[4]

The difficulty index for the items in the test that I am going to construct will be based on the second approach, because it is much easier to use. The ideal difficulty of the items would be at a point on the difficulty scale midway between zero difficulty (100 percent correct response) and chance level difficulty (50 percent correct for true-

false items and 25 percent correct for four-alternative multiple-choice items). Another approach that I am going to use to obtain the required item difficulty is by using the measurement of a discrimination index. The discrimination index (DI) "measures the extent to which the results of an individual item correlate with results from the whole test" (Alderson *et al.*, 1996: 80).[5]

3.3.5 Planning the number of items in the test and its parts

The number of items in the test depends on three main factors: the time allocated for the test ; the level of difficulty of the questions; and the level of students sitting the test. Ebel (1979) is of the opinion that it is difficult to specify precisely how many items should be included in a given test. However, he suggests that :

> "...test constructors might assume that typical multiple-choice items can be answered by even the slower student at the rate of one per minute, and the true-false items can be answered similarly at the rate of two per minute [and] an essay question or a problem depends on the nature of the question or a problem..." (p.78).

Since the time limit for the test is not more than three hours and the students have different levels of ability, the total number of items in the test will be approximately as follows:

Topics / skills:	Item types:	no. of items:	time approx.
(A) Listening to comprehensive passage	multiple choice	15	35 minutes*
(B) Listening & writing	dictation	one passage	20 minutes
(C) Reading	cloze test	two passages	20 minutes
(D) Reading of comprehensive passage	true-false	20	20 minutes**
(E) Reading of comprehensive passage	multiple-choice	10	10 minutes
(F) Writing	short essay	one passage	30 minutes
(G) Arabic grammar	true-false	40	30 minutes
(H) Arabic grammar	multiple-choice	15	10 minutes

* Time allocated for listening to comprehensive passage includes listening to the recorded voice on the tape recorder.

** Approximately ten minutes are allocated for reading passage(s) in the reading to comprehensive passage.

Note: the total time for the test is approximately two hours fifty minutes which allows an extra ten minutes as reserve time for revision, correction, etc.

3.4 The description of the preliminary test

In this section, the discussion focuses first on a general description of the test and then on a detailed description of each test. It should be noted here that the following description is at the preliminary stage only. The final test framework will be settled only after the pre-test (pilot study) has been carried out, and after consultation with experts in this field.

3.4.1 General description

The purpose of the Arabic Placement Test, as stated above, is to measure students' abilities linguistically and communicatively. The battery includes two subject-related tests: a test of listening, reading and writing skills ; and a test of Arabic grammar. The listening test consists of two types of items: multiple choice and dictation. The Reading Test consists of three types of items: cloze, true-false and multiple choice. The Writing Test consists of two types of items; a short essay and dictation which also tests listening skills. The test of Arabic grammar consists of two types of items: true-false and multiple choice. The total number of sub-tests is five. The test does not assess spoken language, since oral tests can only be undertaken on an individual basis. The need in placement testing is essentially for a group test which can be taken in a short period of time.

The materials for the test cover general topics and topics specifically related to the students' areas of study, i.e. *Sharī`a*, *Uṣūluddīn*, and *Tarbiya Islāmiyya*. The rationale for including a specific and a general topic of Arabic is that this will allow for an investigation of the relationship between language competence and background

knowledge. We can thus determine theoretically and empirically which topic better predicts students' future performance, i.e. academic success.

To prepare the test materials, I have referred to text books and lecture notes from the Year One syllabus at the AIS. These materials have been supplied by the AIS for this particular purpose. I have also referred to materials from the departmental library of the Department of Islamic and Middle Eastern Studies at the University of Edinburgh.

The material for the listening test is not available at this stage. It is hoped that these materials can be obtained during a visit to some Arab universities which I will undertake in the near future. I also hope to collect material when I return to Malaysia to administer the preliminary test, including material from lectures at the AIS, some of which will be used as simulated lectures for the tests. In the discussion below, I will explain each sub-test in terms of its content and will give brief illustrations of items.

3.4.2 Test material

The discussion below focuses test materials for the draft sub-tests, which will be used in the pilot study. Four topics will be discussed in this section: item writing, content and description of the test, analysis of behavioral objectives, and methods of scoring. The Reading Test is discussed first, followed by Grammar, Essay and lastly Dictation Tests.

3.4.2.1 Test A: Reading comprehension

3.4.2.1.1 Item writing

In order to prepare the materials for this type of test, I have referred to several resources such as text books for Year One at the AIS, test items from various institutions such as Center for Applied Linguistics at Washington and the Language Centre at King Saud University and the Language Centre at Jordan University. The purpose of investigating these materials is to get a general idea as to what types of reading materials are being used for the purpose of placement or proficiency tests.

3.4.2.1.2 Content and description of the test

The test consists of ninety-nine items divided into three parts: Part One is multiple choice with ten questions; Part Two is true-false type with twenty questions; and Part Three is cloze-test type with sixty-nine questions. Below are the details of the test content followed by the description of the content:

(i) Test content.

As stated earlier in 3.3.1, the materials for the test are taken from general topics as well as from topics specifically related to the students' areas of study, i.e. *Sharī`a, Uṣūluddīn,* and *Tarbiya Islāmiyya*. The rationale for including a specific and a general topic of Arabic is that this will allow for an investigation of the relationship between language competence and background knowledge and thus determine theoretically and empirically which topic better predicts the students' future performance, i.e. academic success.

Part One (see Appendix A.2.1: 440-442): Part One consists of multiple choice questions. It includes four texts ranging from general topics to Islamic religious

matters with different levels of difficulty in terms of vocabulary usage and themes. For example, the first two texts are considered as a test of general Arabic: the first text is about an orator asked on how long he needs to prepare his speeches ranging from short speeches to long ones and the second text is about the harmful effects of being a smoker. The last two texts are about the responsibility as a parent in looking after his or her children from the Islamic point of view. We may assume that these two texts are related to Islamic religious matters. The number of questions for every text can be summarised as follows:

Text:	**No. of questions**
One	3
Two	3
Three	2
Four	2

Part Two (see Appendix A.2.1: 442-444): Part Two is a true-false type of test. It includes five texts followed by twenty questions all together. As in Part One, topics for every text differ in terms of themes, difficulty and vocabulary usage. The relatively challenging topics, for example, are the last two texts. The following is a summary of every text:

Text One: This text is followed by two questions. It concerns the obligatory prayer which is fully related to the students' areas of study, i.e. *Sharī`a*, *Uṣūluddīn*, and *Tarbiya Islāmiyya*. The text itself was taken from the primary reference book for the Faculty of *Sharī`a* at the AIS.

Texts Two and Three: the texts, followed by eight questions, cover general Arabic and are not directly related to the students' area of study. With some alterations, both texts are taken from the Arabic Proficiency Test paper prepared by the Center for Applied Linguistics at Washington.

Text Four: this text is also about general Arabic. However, it is relatively more difficult than the first three texts in terms of vocabulary usage and is followed by five true-false questions.

Text Five: this text , followed by five questions, is again fully related to the syllabus at the AIS. It may be suggested here that if the candidate can answer all the questions from this text correctly, his or her ability to study major subjects at the AIS is very high; the level of difficulty of the text is high and only those who are fluent in Arabic will be able to do it.

Part Three (see Appendix A.2.1: 444-445): this part, which is a cloze test, can be classified as the most difficult task in the reading comprehension test. There are two texts in this part: the first has 24 blanks with the deletion of every fifth word, and the second has 45 blanks with the deletion of every sixth word. The first and the last sentences for each text are left intact to provide lead-in and lead-out context.

In selecting the test content, I have considered several factors such as intellectual content, cultural content, linguistic difficulty, and register and level of formality. Taking, for example, the first text, which is about the holy journey made by the Prophet Mohammed (Peace be upon him) to the Mosque of *al-Aqṣā* and then to bear witness to Allah at the *Sidrat al-Muntahā*. This journey is known as the *al-Isrā' wa'l-mi`rāj*. This important event is well known among Muslims because it is celebrated every year. The second text is also related to Muslim culture. It is about ethics and morality in Islam which have been translated into every-day activities such as eating, drinking, going to the toilet, wearing clothes and meeting people. In other words, the contents of both texts are not new to the examinees. It is important to stress here that the passages are not accompanied with any illustration, diagrams etc.

so that no information other than that provided by the reading texts could be used in allocating the missing words. Also neither text refers to remarks either before or after them.

(ii) Description of the test

The description of the test can be summarised as follows:

Cover page: The Arabic rubric instructs the candidates to write their names and other information on the answer sheet, and also gives the total number of questions, and the total marks for each question. The time allocated is fifty minutes only: ten minutes for Part One, and twenty minutes each for parts two and three. Candidates are not allowed to write their answers on the question papers.

As stated earlier, the test consists of three parts: Part One consists of the multiple-choice questions, Part Two consists of true-false type questions, and Part Three is a cloze test. For multiple choice in Part One, the examinees have to indicate the correct answers by making a circle around one of the four choices on their answer sheet, which is provided separately. The number of correct answers for every choice is as follows: *alif* and *dal*, ahve four correct answer each; *bā'* and *jīm* have one correct answer each. A sample of an answer sheet is provided in the Appendix A.2.5: 496-97)

In Part Two, the examinees are required to determine whether the statements provided in the questions are true or false by marking (✔) for the true statements and (✘) for the false statements. The total number of statements in this part is twenty: eleven are true statements and the remaining nine are false statements. This test differs slightly from the Grammar Test discussed below because it does not ask the examinees to provide the correct answers for the false statements. This is because it

is more difficult for the examinees to write down the true statements. They would in this case be obliged to refer back to the texts, which do not contain exact facts as in the Grammar Test. Moreover, composing original statements can take up too much time and answers may vary between candidates.

In Part Three, the examinees are required to fill in the blank words which are deleted mechanically.

3.4.2.1.3 Analysis of behavioral objectives

The analysis of behavioral objectives was used in designing the questions based on the formula of Taxonomy Bloom. The table below shows the number of items in each category for multiple choice type and true-false type (see Appendix A.2.1: 440-44):

Taxonomy categories:	**no. of items and question nos. (in bracket).:**
1. knowledge	10 (1, 8, 9, 21, 22, 25, 26, 27, 29, 30)
2. comprehension	6 (10, 16, 17, 18, 24, 28)
3. application	nil
4. analysis	7 (4, 7, 13, 14, 15, 19, 23)
5. synthesis	2 (2, 3)
6. evaluation	5 (5, 6, 11, 12, 20)

3.4.2.1.4 Methods of scoring

There are basically two types of marking of the whole test in this experiment: objective and subjective. Objective marking is used for multiple choice, true-false, error-recognition, and other item types where the candidates are required to produce a response which can be marked as either 'correct' or 'incorrect'. Subjective marking usually refers to assessing tests of writing or speaking in which the examiners make

judgements using a rating scale. The scale may consist of numbers, letters or other labels eg. 'Excellent' or 'Poor' "...which may be accompanied by statements of the kind of behaviour that each point on the scale refer to" (Alderson *et al.* 1996: 107). For rating essays alone, there are four principle types of scoring scales: holistic, analytic, primary trait, and multitrait (Cohen, 1994).

The full set of acceptable answers is called a 'key' or 'mark scheme', depending on how much need there is for examiners to exercise their discretion in marking (Mathews, 1985). The methods of scoring for each of these will be described below in accordance to what format each sub-test uses.

For the Reading Test, two methods of scoring will be used:

(i) For multiple choice format, the *key* which is the correct answer to every item will be provided to the examiners to calculate the total marks obtained by every candidate. In addition, each response (a, b, c, d) will be given one point: 1 for (a), 2 for (b), 3 for (c), and 4 for (d). Zero will be given for unanswered questions. The reason for giving numbers at this stage is because every answer whether it is right or wrong will be analysed to obtain an item facility for every item of the test which will be discussed later.

(ii) For true-false and cloze test formats, the *mark scheme* method is used because there is more than one possible response for an item. Therefore, every incorrect response will be marked with number one, every correct response will be marked with number two, and every unanswered question will be marked with zero. Since the ensuring of the correct answer in the cloze part is more flexible, i.e. any contextually appropriate replacement such as something that is grammatically appropriate, semantically appropriate or both together will be implemented in the

real test, the examiners are required to make a record of unpredicted answers that are acceptable.

3.4.2.2 Test B: Arabic grammar

3.4.2.2.1 Item writing

Generally a criterion-referenced approach to item writing is used. This means the items test students in terms of specific performance, i.e. what an individual can do, without reference to the performance of other candidates (Gronlund, 1982). Items are constructed each measuring a specific aspect of the grammar topics assigned to Year One syllabus at the AIS. To prepare the items, I have analysed the Arabic grammar textbook written by language teachers at the Faculty of Language and Linguistics at the University in an attempt to identify those grammatical elements most likely to be used by the students when dealing with Arabic and with subject related materials, i.e. their major courses.

In addition to the analysis of the text book, I have also referred to the findings of a published research project conducted by myself (Dahan, 1996) on error analysis in writing done by final year students at the AIS. In my study, I found that some mistakes are related to Year One Arabic Syntax such as indefinite and definite forms, nominal sentences, and subject and predicate; and for morphology, the mistakes made by the samples relate to singular, dual, and plural, aplastic and inflected nouns and verbs, and verbs that are bare of any accessory and verbs that have an accessory. Although the findings of the research have no direct relation to this study, these findings are useful in so far as I am dealing with a homogeneous population who share the same first language. In this regard, Lado (1961) is of the view that since the non-

native speakers have a mother tongue which differs from the target language, in this case Arabic, we need to identify areas of differences using contrastive analysis and then test them.

3.4.2.2.2 Content and description of the test

The test consists of 65 items divided into two types: multiple choice and true-false. These items represent the grammatical topics for the Year One syllabus at the AIS. Below are the details of the test content followed by the description of the test:

(i) Test content (see Appendix A.2.1: 446-51):

The particular topics of grammatical structure and the total number of items which are included in the composition of the test are:

Topic:	**no. of items**
1. Syntax	36
2. Morphology	14

The details of the syntax topic covered by the multiple choice are as follows:

declension and indeclension	6
indefinite and definite	6
subject and predicate	6
inna and its sisters	10
kana and its sisters	8
Total	36

The details of the morphology topic covered by the multiple choice are as follows:

singular, dual, and plural	3
aplastic and derivative nouns	4
aplastic and inflected noun	7
Total	14

The total number of items, according to topics for true-false type is as follows:

1. Syntax 7
2. Morphology 8

Topics for syntax and morphology covered in the true-false type are relatively similar to those covered in the multiple choice.

(ii) Description of the test

The description of the test can be summarised as follows:

Cover page: The Arabic rubric instructs the candidates to write their names and other information on the answer sheet, and also gives the total number of questions, and the total marks of each question. The time allocated is forty minutes only: thirty minutes for Part One and ten minutes for Part Two. Candidates are not allowed to write their answers on the question papers.

The test consists of two parts: Part One consists of multiple-choice questions and Part Two is the true-false questions. Both parts provide an example on how to answer the questions.

Part One (see Appendix A.2.1: 446-50): With the stem, the multiple choice questions ask or imply a direct or an indirect question to acquaint the examinee with the problem that is being posed. (to view an example of direct and indirect questions, see items no. 13 and 4 respectively in Appendix A.2.1: 446, 447):

Four responses are provided: *alif*; *bā'*; *jīm*; and *dāl*. Since multiple-choice responses are all intended to be answers to the same question, I have chosen the three distractors as much as possible to be parallel in grammatical structure, in type of content, in length, and in complexity with the correct answer. Taking as an example question no. 5: the four choices use the same word *akh* (brother) as a root and the

same pronoun *nā* (us) attached to that word. The only differences are the distractor (*alif*) in the genitive form; the distracter (*bā'*) in the accusative form; and the distracter (*jīm*) in the plural form. The answer, which is (*dāl*), is in the nominative form.

The examinees have to indicate the correct answers by making a circle around one of the four choices on their answer sheet which is provided separately. The proportion of the total correct answer for every choice is as follows: *alif* has eleven correct answers; *bā'* has fourteen correct answers; *jīm* has twelve correct answers; and *dāl* has thirteen correct answers. A sample of an answer sheet is provided in Appendix A.2.5: 498, 499.

Part Two (see Appendix A.2.1: 450-51): The examinees are required to determine whether the statements provided in the questions are true or false by marking (✔) for the true statements and (✘) for the false statements. Fifteen statements have been derived, six of them false. In addition to marking, the examinees have to provide the correct answers for the false statements they mark. The risk is that correct answers may be achieved by guess-work.

3.4.2.2.3 Analysis of behavioral objectives

The analysis of behavioral objectives has been used in designing the questions based on the formula of Taxonomy Categories[6]. Most of the questions in Part One which are the multiple choice can be characterised as knowledge and application categories. For example, in the knowledge category, the examinees are required to name, or to match the answer with what they have learned of Arabic grammar. On the other hand, in the application category, the examinees are required to demonstrate

their understanding of Arabic grammar when dealing with this type of question. Question nos. 1, 2, 3, and the like are the sample for the application category while question nos. 13, 15, 18, and the like are the sample for the knowledge category.

The true-false questions in Part Two have been characterised as the synthesis category. The examinees are required to combine their knowledge of Arabic grammar with the statements laid in the questions in order to determine whether those statements are true or false. The table below shows the summary of the taxonomy categories for the Arabic Grammar paper:

Taxonomy categories:	**no. of questions**
1. knowledge	17
2. comprehension	0
3. application	33
4. analysis	0
5. synthesis	15
6. evaluation	0
Total	**65**

From the above table, it may be argued that there is no balance between the taxonomy categories used in the Arabic Grammar Test. It is important to stress here that this does not mean that there should be a uniform distribution of the taxonomy categories in the table of specifications. Rather, the balance should depend on the subjects and areas of study. Arabic grammar, for example, deals with facts rather than with opinions. Therefore it is difficult to design questions which can be characterised as comprehension, analysis, and evaluation categories.

3.4.2.2.4 Method of scoring

The Grammar Test has two formats: multiple choices and true-false which can be classified for the scoring method under objective marking. Two methods of scoring have been used to mark students' response to this test.

(i) For multiple choice format, the *key* which is the correct answer to every item will be provided to the examiners to calculate the total marks obtained by every candidate. In addition, each response (a, b, c, d) will be given one point: 1 for (a), 2 for (b), 3 for (c), and 4 for (d). Zero will be given for unanswered questions. The reason for giving numbers at this stage is the same what has been mentioned above in the Reading Test; and,

(ii) for true-false format, the *mark scheme* method is used because there is more than one possible response for an item. Therefore, every incorrect response will be marked with number one, every correct response will be marked with number two, every half-correct answer will be marked with number three, and every unanswered question will be marked with zero.

3.4.2.3 Test C: Essay

3.4.2.3.1 Item writing

Although the Dictation Test involves certain writing skills, students also need to be tested on this skill in more depth, i.e. producing an original piece of writing in Arabic for formal academic writing tasks. The topic of the essay is: A New Life at

University. Some major points are provided in the question paper to help the examinees to write down their essay.

3.4.2.3.2 Content and description of the test

(i) Test content (see Appendix A.2.1: 452)

As stated above, the topic is very relevant to the examinees because it reflects their own experience. The writing starts with the first point which is about getting the result of the final examination at secondary school or at pre-university level. All of the examinees have gone through this stage otherwise they would not be in the examination hall on that day. The examinees are required to write about their feelings and reactions to the result of the final examination. The second point highlights the procedures that the examinees have gone through when applying for a place at the university where they are now. These procedures involve requesting an application form, filling in the form, and then sending it to the respective body so that the form can be processed. The third point focuses on the happiest events in the examinees' academic life: receiving an offer of admission to the university. They are expected to express their feelings concerning this particular event, and then to describe the preparations for going to the university which might involve buying clothes, books, tickets, and preparing documents such as certificates, photos, x-ray film, medical reports, etc. The last point focuses on the most exciting part of the whole experience that is, the first day at university. The examinees are expected to differentiate between life at secondary school and life at university, how the seniors welcome them, the quality of the food in the halls of residence, and how big the library, lecture halls, and gymnasium are, etc. It should be clear from the above that the examinees should

not have any difficulty in writing on the suggested topic since it touches every one of them.

(ii) Description of the test

Cover page: The Arabic rubric instructs the candidates to write their names, their answer and other information on the answer sheet, and also the total marks of the question, and that the time allocated is thirty minutes only. Candidates are not allowed to write their answers on the question papers. The duration of the test is 30 minutes. The examinees are required to write the essay on the provided answer papers only and no maximum or minimum number of words is required. With regards to format, the examinees have a choice whether to write their essay in letter format or descriptive or dialogue formats or any other format they are interested in, using the main points given in the question paper. They are allowed also to add any points they think relevant to the topic of the essay since the purpose of listing major points in the question paper is to help them to gather the ideas of the topic only.

3.4.2.3.3 Analysis of behavioral objectives

The analysis of behavioral objectives has been used in designing the questions based on the formula of Taxonomy Categories. Some of the taxonomy categories will be employed in this test. For example, in the knowledge category, the examinees have to use their knowledge of specifics to choose the right terminology and facts in their writing. In the application category, the examinees have to demonstrate their understanding of Arabic grammar that includes syntax and morphology when dealing with this type of exercise.

3.4.2.3.4 Method of scoring

Subjective marking that refers to assessing tests of writing or speaking in which the examiners make judgements using a rating scale is employed in this test. Among the four types of scoring scales, namely holistic, analytic, primary trait, and multitrait, I will choose the analytic scale. According to Weir (1990), there are two main advantages in using analytic scoring: analytic scales guard against the collapsing of categories; and, training of raters is easier when there is an explicit set of analytic scales.

Each scale in the analytic scoring will assess different aspects of writing. Four aspects of writing will be assessed. They are content and organisation (10 marks), vocabulary (5 marks), grammar (5 marks), and mechanics (5 marks) which bring the total of 25 marks.

Below are the details of the scales being used in giving marks to every aspect of the writing that is being assessed (taken from Cohen 1994: 328-29):

1. *Content and organisation*

Marks:	**Label**	**Criteria**
9-10	Excellent	main ideas stated clearly and accurately, change of opinion very clear; well organised and perfectly coherent
7-8	Good	main ideas stated fairly clearly and accurately, change of opinion relatively very clear;
5-6	Average	main ideas somewhat unclear or inaccurate, change of opinion statement weak; loosely organised but main ideas clear, logical but incomplete sequencing
3-4	Poor	main ideas not clear or accurate, change of opinion statement weak; ideas disconnected lacks logical sequencing
1-2	Very poor	main ideas not at all clear or accurate, change of opinion statement very weak; no organisation, incoherent

2. Vocabulary

Marks:	Label	Criteria
5	Excellent	very effective choice of words and use of idioms and word forms
4	Good	effective choice of words and use of idioms and word forms
3	Average	adequate choice of words but some misuse of vocabulary
2	Poor	limited range, confused use of words, idioms, and word forms
1	Very poor	very limited range, very poor knowledge of words, idioms, and word forms

3. Grammar

Marks:	Label	Criteria
5	Excellent	no errors, full control of complex structure
4	Good	almost no errors, good control of structure
3	Average	some errors, fair control of structure
2	Poor	many errors, poor control of structure
1	Very poor	dominated by errors, no control of structure

4. Mechanics

Marks:	Label	Criteria
5	Excellent	mastery of spelling and punctuation
4	Good	few errors in spelling and punctuation
3	Average	fair number of spelling and punctuation errors
2	Poor	frequent errors in spelling and punctuation
1	Very poor	no control over spelling and punctuation

3.4.2.4 Test D: Dictation

3.4.2.4.1 Item writing

This test is aimed at assessing students' ability to listen, understand, and then write information dictated to them. This is very important for their every-day academic life because the Dictation Test in a way resembles the task of note-taking in a lecture room situation. The only difference is that in the Dictation Test the examinees are required to write what they hear in full. The passage to be dictated has

been prepared by me and to the best of my knowledge, it does not contain highly complicated words. The theme of the text is very familiar to the examinees. It is about preventing people from doing bad things using three methods suggested by the Prophet *Moḥammed* (peace be upon him) in his famous *Ḥadīth*. The purpose of selecting a theme that is familiar to the examinees is to ensure that they will not have to deal with a dictation topic that is totally new to them. If this principle is not observed, the Dictation Test may prove to be very difficult, especially for beginners at the university level.

3.4.2.4.2 The description of the test (see Appendix C: 546 for the recorded voice)

The entire Dictation Test - including instructions, the introductory passage, and the test passage - is written by me and is recorded on audio tape by a native speaker[7]. This is the only test in which the examinees are not provided with question papers. They are provided only with the answer sheet to write down what is read to them. The duration of the test is approximately 25 minutes. This includes:

(a) instructions for the test;

(b) the first reading of the test passage during which the entire text is read to the examinees at normal conversational speed while they listen only;

(c) the second reading of the test passage which includes pauses (approximately 39 pauses) after each of the test's segments to allow the examinees to write down what they hear;

(d) the third reading of the test passage in which the text is read at the same speed as the first reading to allow the examinees to check and correct their work, if necessary; and

(e) a two minute pause for a final check and corrections.

I have allocated in the test passage the length of pauses (in seconds) by writing down the relevant number of seconds in brackets after each segment. This length is set by spelling letter-by-letter each word sequence once, before proceeding to the next segment in the passage. The minimum length of each segment is one word and the maximum is five words. The length of the pauses varies between three and fifteen seconds. (The full texts including an answer sheet are attached in Appendix A.2.1: 453 and A.2.5: 504-5).

In the test instructions, the examinees are reminded not to write down the marks of punctuation, e.g. comma, full stop, semi-colon, etc, in the test passage but to write them down exactly as they are. During the second reading, the marks of punctuation are mentioned again together with the test passage so that the examinees can put them down on their answer sheet. Whitaker (1982) recommends excluding the announcement of the marks of punctuation because they are not part of the spoken language and suggests letting the examinees judge by themselves where to supply punctuation marks. However, I find that mentioning these marks helps the examinees to understand the passage better.

3.4.2.4.3 Analysis of behavioral objectives

The analysis of behavioral objectives has been used in designing the questions based on the formula of Taxonomy Categories. The knowledge category has been used because candidates have to identify and to select which word to write when they listen to the text. Some elements in the application category have been used too because candidates have to demonstrate their ability to choose the right word for their

writing. Therefore it can be said that the Dictation Test combines two types of skill in one test: the ability to receive what is spoken, i.e. listening skills, and the ability to produce what is heard in writing, i.e. writing skills.

3.4.2.4.4 Method of scoring

Objective marking is used for the Dictation Test because the candidates are required to produce a response that can be marked as either 'correct' or 'incorrect'. From the text, I have selected 27 words and items of punctuation that will be given a mark if candidates write them correctly. The selection of these words and punctuation is based primarily on the error analysis of non-native speakers of Arabic that was conducted in the AIS in 1994 (Dahan, 1996). Below are the details of words and punctuation that have been chosen for the marking scheme:

Particulars:	**Question nos.**
1. combination words:	
(i) two words that are pronounced as one word	9, 11, 20
(ii) three words that are pronounced as one word	19
(iii) four words that are pronounced as one word	1, 24
2. the use of *alif lam qamariyya*	3, 5, 10, 15.
3. the use of *alif lam shamsiyya*	4, 12,
4. the use of *iḍāfa*	8, 23
5. the use of long and short vowels	18, 26, 27.
6. the use of *hamzat al-waṣl*	15, 23, 24
7. the use of certain *ḥurūf*:	
(a) *ḥarf Qāf*	7
(b) *ḥarf dhāl*	13
(c) *ḥarf ḍād*	14

	(d) *ḥarf ẓā'*	17
	(e) *ḥarf ṭā'*	22
	(f) *ḥarf dāl*	25
8.	the use of punctuation:	
	(a) colon	6
	(b) comma	16
	(c) full stop	21
9.	the use of *fi`l al-majzūm*	2

3.5 Summary of Chapter Three

In this chapter, I have designed the test specification as a guideline for constructing a draft of the Arabic language placement test, for the purpose of this research. The design of the test specification was not an easy task. Therefore, I have discussed, prior to the establishment of the test specification, the various steps to be followed. These steps, as discussed in detail above, provide very important information as well as giving a general guide to ensure that the establishment of the test specification reflects the characteristics of the test. After establishing the test specification, one set of the draft test was set up. Four main aspects of the construction of this draft test were discussed. They are item writing, the content and description of the test, the analysis of behavioural objectives, and the methods of scoring. It is hoped that the draft test, the construction of which is based on the test specification, will have a high face and content validity. These two types of validity will be discussed in the next chapter.

End notes:

[1] Some other references which are important for preparing test specifications can be found in Gronlund (1982), Alderson *et al.* (1996), and Bachman and Palmer (1996).

[2] A classification of educational objectives. Adapted from Bloom (1956: 201-207)

A. *Knowledge*

1.00 KNOWLEDGE
- 1.10 Knowledge of Specifics
  - 1.11 Knowledge of Terminology
  - 1.12 Knowledge of Specific Facts
- 1.20 Knowledge of Ways and Means of Dealing with Specifics
  - 1.21 Knowledge of Conventions
  - 1.22 Knowledge of Trends and Sequences
  - 1.23 Knowledge of Classifications and Categories
  - 1.24 Knowledge of Criteria
  - 1.25 Knowledge of Methodology
- 1.30 Knowledge of Universals and Abstractions in A Field
  - 1.31 Knowledge of Principles and generalisations
  - 1.32 Knowledge of Theories and Structures

B. *Intellectual Abilities Skills*

2.00 COMPREHENSION (Grasping the meaning of material)
- 2.10 Translation (Converting from one form to another)
- 2.20 Interpretation (Explaining or summarising material)
- 2.30 Extrapolation (Extending the meaning beyond the data)

3.0 APPLICATION (Using information and abstractions in a particular situation)

4.00 ANALYSIS (Breaking down material into its parts)
- 4.10 Analysis of elements (Identifying the parts)
- 4.20 Analysis of relationships (Identifying the relationship)
- 4.30 Analysis of organisational principles (Identifying the organisations are arranged)

5.00 SYNTHESIS (Putting parts together into a whole)
- 5.10 Production of a unique communication
- 5.20 Production of a plan or proposed of operations
- 5.30 Derivation of a set of abstract relations

6.00 EVALUATION (Judging the value of a thing for a given purpose using definite criteria)

6.10 Judgement from internal evidence

6.20 Judgement from external criteria

[3]. An illustration of taxonomy categories with Action verbs by Gronlund (op. cit., 23)

TAXONOMY CATEGORIES	SAMPLE VERBS FOR STATING SPECIFIC LEARNING OUTCOMES
KNOWLEDGE	Identifies, names, defines, describes, lists, matches, selects, outlines
COMPREHENSION	Classifies, explains, summarises, converts, predicts, distinguishes between
APPLICATION	Demonstrates, computes, solves, modifies, arranges, operates, relates
ANALYSIS	Differentiates, diagrams, estimates, separates, infers, orders, subdivides
SYNTHESIS	Combines, creates, formulates, designs, composes, constructs, rearranges, revises
EVALUATION	Judges, criticises, compares, justifies, concludes, discriminates, supports

[4] The difficulty index may vary from $p = 0$ for an item answered correctly by no examinee to $p = 1.00$, for an item answered correctly by all examinees. The difficulty index may be adjusted to remove the chance success by the examinees (Tinkelman, 1976). The chance success always happens in a question that has multiple answers. In a multiple-choice item or true-false items, it is likely that a certain percentage of the examinees who do not know the answer correctly answer the item as a result of guessing. Even though the extent to which guessing can never be precisely determined in the case of any specific item, it is often useful to estimate what percentage of the examinees actually knew the correct answer. Tinkelman (op. cit., 62-3) suggests the following to eliminate the effect of chance success by the examinees:

> "For example, if a four-option item is answered correctly by 70 percent of the examinees, then it can be assumed that the 30 percent who failed the item and apparently guessed wrong were distributed, on the avarage, with 10 percent guessing each of the three wrong alternatives. Most probably, therefore, there were another 10 percent who guessed the correct answer, so that the percentage of candidates who actually knew the correct answer was very likely closer to 60

percent than 70 percent."

[5] How to obtain the Discrimination index (Alderson *et al.* 1996: 274)

1. Rank the students according to their total score.
2. Devide them into three groups, making sure that the top and bottom groups have equal numbers of students.
3. Count how many students in the top group get an item right, and how many in the bottom group.
4. Find the different between the number of correct answers in the top group (RT) and the number of correct answers in the bottom group (RB). Divide this by the total number of people in the top group (NT):

$$\frac{RT - RB}{NT}$$

[6] See end note no. 3 for the detailed explanation about the Taxonomy Categories.

[7] He is Mr. J. Giaber, a postgraduate student from Libya who is currently studying for a Ph.D degree in translation at the University of Edinburgh.

4. CHAPTER FOUR: PILOT EXPERIMENT AND INTERNAL VALIDITY

4.1 Introduction

In Chapter Three, I discussed the test specifications and the procedure for writing test items. One set of the preliminary test booklet was prepared at the end of the chapter. The next step which will be presented in this chapter is the pilot survey of the preliminary test items, the collection of data as a basis for improving items and selecting the best available items to form the final test. Two sub-topics will be discussed in this chapter: the pilot experiment of the draft test, and the investigation of the internal validity of test items using the tools of statistics: *descriptive statistics* and *item analysis*. To prepare the ground for carrying out this task, this chapter will discuss first the fundamental literature of the pilot test administration. This includes the purpose of the pilot test, stages involved in the pilot test administration, rules and regulations for conducting pilot test etc. This is followed by a brief explanation of the instruments of statistics employed in the analysis. This involves descriptive analysis, which involves *central tendency* and *dispersion*, and item analysis, which involves *item facility* (IF), *item discrimination* (ID) and *distractor efficiency* (DE) analyses.

4.2 Pilot testing in the testing administration

The word pilot or pretest or tryout refers to "...all trials of an examination which take place before it is launched or becomes operational or 'live'..." (Alderson *et al.* 1996:74). The main purpose of such a test is to collect information about test

usefulness in order to make revisions in the test itself, rather than to make inferences about individuals (Bachman and Palmer, 1996). In this regard, Alderson *et al.* (op. cit: 73) are of the view that it is very important to do the pilot experiment and try out the items:

> "However well designed an examination may be, and however carefully it has been edited, it is not possible to know how it will work until it has been tried out on students. Although item writers may think they know what an item is testing and what the correct answer is, they cannot anticipate the responses of learners at different levels of language ability."

Therefore, in order to select the best items from the draft test and to make improvements in weaker items and in order to drop the weakest items, I need to run one or more pilot surveys.

4.3 Purpose of pilot study

According to Henrysson (1971), the pilot surveys provide data for such purposes as:

1. Identifying weak or defective items. For example, from the tryout, we can identify ambiguous or indeterminate items with nonfunctioning or implausible distracters;
2. Determining the difficulty of each item so that a screening may be made in order to have a distribution of item difficulties appropriate to the purpose of the final test using item facility or item difficulty;
3. Determining for each item its power to discriminate between good and poor

students in the achievement variable being measured using the item discrimination index;

4. Determining how many items should constitute the final test;
5. Determining appropriate time limits for the final test;
6. Discovering weaknesses in the directions to examinee and to examiner, in the sample or practice exercises, in the format, and so forth;
7. Determining the intercorrelations among the items to avoid too much overlap or bias in item selection and to check the grouping of items into sub-tests.

4.4 Stages of pilot survey

The importance assigned to any one of the above purposes and the nature of the pilot survey will vary with the type of test and the amount of time and resources available. Henrysson (op. cit.) views that the whole pilot procedure can ideally be divided into three stages: pretryout or pretesting, tryout or pilot study, and trial administration of the final test.

1. Pretryout means a preliminary administration of test items to a small sample of students from the population on which the test is to be used. At this stage, the procedure may be informal and may involve only the administration of a mimeographed set of items. A test constructor does not expect to make a complete statistical item analysis of the data collected. Henrysson suggests that the test constructor may wish to administer the pretryout himself since much may be learned by direct observation and personal interview. Alderson *et al.* (1996) however advise that for the first stage at least two of the samples should be native

speakers of the language being tested. The purpose of this is to enable them to see whether the instructions are clear, the language of the items acceptable and the answer key accurate. Alderson *et.al* (op. cit.) argue that if this is not observed, faults may emerge in a test "...especially if the test constructors do not have the language being tested as their first language" (p.74).

2. After a pretryout has been completed and most of the gross deficiencies eliminated, a formal tryout is conducted to obtain more accurate information on each item (Henrysson, op. cit.). Henrysson adds that if there are too many deficiencies revealed in the tryout which call for extensive revisions to the test items, a second full tryout is needed. This is more often the case, according to him, when a new kind of test is being constructed.
3. This stage is conducted based on the data obtained in the second stage, i.e. the tryout. The main purpose of the third stage is to ascertain exactly how the test will function in actual use and to estimate the norms, validity, reliability, etc., of the final test. Henrysson (op. cit.) stresses that no material changes should be made after the trial administration.

4.5 Rules for administering the pilot test

There are rules for administering the pilot study which are similar to the rules for administering the real test. In the discussion of this section, I will restrict the discussion to some points that are specially important for the pilot study.

(i) Sampling

To ensure the sample for the pilot study represents the population in the real

test, I will do my best to select samples representing a similar background and level to those who will take the final version of the test. In this connection, Henrysson (op. cit:132) suggests:

> "Ideally, each student in the sample should be individually drawn from the population by simple random or stratified sampling. However, such a procedure is usually not practical, so some procedure of cluster sampling is used."

According to Alderson *et al.* (op. cit.), one of the main questions facing any test constructor is the number of candidates on whom a test should be trialled. Henrysson suggests that "...300 or more students will be needed" (p.131). He adds that if the final test is to be constructed for use on several age or grade levels, separate samples are needed for each level. Alderson *et. al* however argue that it is impossible to give a rule for this as "...the number depends on the importance and type of exam, and also the availability of suitable students" (p. 75). For example, the construction of traditional multiple-choice items is very difficult and it is easy for the item constructor to miss ambiguities in the distractors. It may be necessary that such items need more pilot testing than any other type (Alderson *et al.* op. cit.). With regards to the number of candidates, Henning (1987) recommends 1000 candidates for trial multiple-choice tests. Alderson *et al.* again argue that it is very difficult to find samples with that number for trialing where sometimes "...test constructors may have to be content with a sample of 200 or 300, or even 30 or 40" (p. 75). Alderson *et al.* conclude that large numbers of samples are not necessary because the main purpose is to achieve invaluable information about the ease of administering the test, the time students need for completing it and so forth. The only guiding rule is the

more the better, since the more students there are, the less effect chance will have on the result.

(ii) Testing conditions

Test constructors should review and revise all the directions for examiners and examinees on the basis of the pilot study. Henrysson (op. cit.) suggests that provision should always be made to secure complete reports and comments from the examiners on any problems and weaknesses arising during the pilot testing. He also suggests that feedback which includes views and comments regarding the directions and the test items from the examinees' point of view in the forms of interview or writing format is not less important. Bachman and Palmer (1996) share the same point of view with Henrysson regarding this matter. They think that this feedback will help test constructors to collect information as much as possible as to what modifications might be required in order to improve the usefulness of the test. Both Bachman and Palmer (op. cit.) add that among the types of feedback test constructors may need are:

(a) Feedback about test takers' language ability

This includes information on the extent to which the test tasks require the test takers to use components of language ability (organisational and pragmatic knowledge) and topical knowledge. "This kind of information is useful in making a preliminary assessment of the construct validity, authenticity, and interactiveness of the test tasks" (p. 238).

(b) Feedback about the testing procedure itself

This involves information on circumstances and events taking place during the trial administration either to activities of the test takers or to activities surrounding the test takers. For example, if in writing a composition test, a number of test takers ask

about the length of the composition to be composed, and so forth, the test constructor may need to revise the instructions, perhaps by providing more information in the instruction. Or, if a test involves using an audio tape and half of the candidates do not hear the tape clearly, more loudspeakers may be needed in the future for the final version of the test.

Furthermore, teachers and those who administer the final operational test should administer at least one of the tryouts (Henrysson, op. cit.). This will help test administrators to understand the nature of the test they will be conducting and the importance of their own role towards the administration of the test. According to Alderson *et al.*(op. cit.), administrators not only need to administer one of the tryouts, but also need to undergo training because "...they are responsible for seeing that the conditions in which the test is given provide all candidates with the best chance possible to display the abilities which are being tested" (p.115). " They may have insufficient experience in answering procedural questions, may not be sufficiently supportive of test takers, may be unable to speak the test takers' native language, and so on" (Bachman and Palmer, 1996:236). Therefore test administrators need to be provided with more background information, simulated training sessions in which they are given feedback on their non-verbal communication, debriefing sessions following operational test use, and so on.

Another important issue in the testing conditions is that the samples involved in the pilot test need to be reminded to take the test seriously and do it as well as possible. Otherwise, the ensuing results may invalidate the whole trialling procedure. One way to ensure the samples take the test seriously is to give trial items to candidates while they are sitting live examinations. When trial items are inserted into

the exam, the candidates will not only be of the appropriate level and background, but will also take the items with the seriousness which is often lacking in trials (Alderson *et al.* op. cit.). Some testers however are worried about giving candidates untried items which might be unclear or extremely difficult and therefore cause anxiety to the candidates. One way of overcoming this is by telling the candidates that some of the items are for trial and will not be marked. One of the examination boards that practices this method is the Princeton Examination Board which conducts the General Record Examination (GRE) papers.

It is also important that the test constructors should inform teachers and those who administer the test, as well as those students who are participating in the tryout, of the scores of the test. This is essential to ensure their continuing cooperation in the future (Henrysson op. cit.). Henrysson stresses that "...if for some reason test scores cannot be reported, an explanation should be provided" (p.132).

(iii) Ensuring adequate tryout of all items

It is extremely important that all of the test items are answered by the examinees in order to gather accurate data about each individual test item (Henrysson op. cit.; Alderson *et al.* op. cit.). To make this happen, it may be necessary for test constructors, at this stage, to be relatively liberal in determining the timing of the exam. Most tests of educational achievement are not conducted at high speed but are administered with liberal time limits so as to place the major emphasis on *level* and *power* rather than on *rate* of work (Henrysson op. cit.). Henrysson (op. cit.) goes on to suggest that the best procedure for the tryout is to provide very liberal time limits so that candidates are given ample time to complete the exam. However, a real problem regarding the time limit arises when items to be tried out are for a speeded

test. If the tryout is conducted exactly according to what is intended in the final test, examinees may not be able to answer the items toward the end, and this will affect the data analysis of the items. On the other hand, if the tryout is conducted with generous time limits, it may be contrary to the procedure of the exam itself. In addition, the mental set and rate of work of the examinees may not reflect the real mental set of those who will be sitting for the real test. Therefore the findings of the tryout may not be useful.

Mollenkopf (1950) and Aiken (1964) have demonstrated that test time limit and item placement can have undesirable effects upon the estimates of item parameters for items appearing late in a tryout test. Three methods can be employed to avoid these problems relating to time limit (Henrysson op. cit: 133):

> "Under one method, the items to be tried out under speeded conditions are followed in the test booklet by a set of "cushion" items, of the same general type as others, that are neither analyzed nor scored. These items should be relatively difficult and time consuming since their only purpose is to keep the faster or abler examinees occupied during the latter part of the test period.
>
> Under the second method, the items to be tried out are placed in different order in two or more booklets: thus if 30 items are to be tried out, items 11-20 in booklet A appear as items 1-10 in booklet B, and item 21-30 in booklet A appear as items 1-10 in booklet C. Thus, every item appears among the first 10 in at least one of the tryout booklets, among the second 10 in one, etc. Item analysis data then is computed for the first 66.6 percent (20 items) in each booklet. The second method is essentially the same as the first - in this case, the cushion items in each form are some of the items being tried out...
>
> A third method, applicable when items designed for parallel forms of tests are to be tried out, consists of interspersing new items in current final forms. The pool of new items is divided into groups and tried out together with the current test on different samples. Sometimes a whole page or section containing only new test items can be included in the test booklet being used in the regular problem.

That the scores on the current test can serve as a criterion variable when validating the new items enhances the value of this method"

Henrysson (op. cit.) adds that if the third method is used, some precautions should be observed. For example, examinees should be informed of the tryout materials in the exam paper to reduce their anxiety. In a test where time limit is very important, the test constructors are advised not to let the tryout items waste the examinee's time with unusual length or great difficulty. Henrysson suggests that this can be avoided by giving the tryout items a separately timed section.

(iv) Surplus items for discard

This is another rule which needs to be observed seriously by test constructors. According to Henrysson (op. cit.), it is very important to tryout more items than are needed for the final test in order to counterbalance the effects of sampling fluctuations in the final test. Henrysson points out that there is no strict rule defining the margin for discard that must be followed by test constructors. However, he suggests that a margin for discard of 20 percent can be assumed as more than adequate. Henrysson (op. cit.) however allows that "...a smaller margin for discard is adequate when the test outline calls for a considerably larger number of items "... because the effect of chance will be correspondingly reduced" (p. 134).

Another factor that calls for preparing surplus items in the tryout is the degree of control desired on item difficulty. If a test constructor has set up a very rigid specification for the distribution of item difficulties in the final test, one must expect some loss of items because these items do not match the specification with respect to difficulty (Henrysson op. cit.). This is not to mention other factors like an assessment

by test specialists of the items which may necessitate removing some items from the test papers.

4.6 The instruments used for the analysis

4.6.1 Descriptive statistics

Descriptive statistics which are numerical representations of how a group of students perform on a test refer to the *central tendency* and *dispersion*. The central tendency describes the typical behaviour of a group in a test and shows how the scores cluster together (Brown, 1988, 1996; Alderson *et al.* 1996). Four statistics are used for estimating central tendency: the mean, the mode, the median, and the midpoint. The mean is derived by adding up all the individual scores of a distribution and dividing the total by the total number of scores in the distribution. The median refers to a point below which 50% of the scores fall and above which 50% fall. The mode tells us about the scores that occur most frequently. In other words, it is a score gained by the largest number of students. Brown (1996) calls the mode as the most 'fashionable score' because it is achieved by the majority of candidates. The last statistical measure for the central tendency is the midpoint. The midpoint in a set of scores is that point halfway between the highest score and the lowest (Brown op. cit.). The dispersion shows how the individual scores are spread out around the central tendency. Three indicators are commonly used for describing the dispersion: the standard deviation (SD), the variance, and the range. The SD is, "...approximately, the average amount that each student's score deviates (differs) from the mean" (Alderson, *et al.* op. cit: 95). The SD provides us with a more

accurate description of the spread of the scores. If a distribution were perfectly normal, we would expect to fit exactly 3 SDs on either side of the mean: the first on either side of the mean would account for 68%, the second for 95% and the third for 99.7% of the population under analysis (Green and Weir, 1998). The variance can be defined as the average of the squared differences of candidates' scores from the mean (Brown, 1996). Test variance is also known as "...the square of the standard deviation, or as an intermediary step in the calculation of the standard deviation" (op. cit. p. 109). The range is the number of points between the highest score and the lowest score. The range provides some idea of how individuals vary from the central tendency. Some writers ignore the range in their descriptive statistical report because the range only reflects the magnitude of the outer edges of all the variation in scores and therefore is considered a weak measure of dispersion (see Alderson *et al.* 1995). Some factors like personal decisions of test takers or anything unusual that happens to the candidates during the test event can strongly affect the range even though they are extraneous to the candidates' performances of the test. Regardless of this problem, the range can still be beneficial because it is the point between the highest and lowest scores of a test.

4.6.2 Item analysis

Three instruments will be used to develop the items of the test, namely item facility (IF), item discrimination (ID), and distractor efficiency (DE) analyses. IF formula, which is also called item difficulty or item easiness, is useful to examine the percentage of samples who correctly answer a given item. Using this formula, the result will be arranged ranging from 0.00 to 1.00 value. This value can be interpreted

as the percentage of correct answers for a given item by moving the decimal point two places to the right (Brown, 1996). For example, if the index of IF is .16, this means that only 16% of the samples answered the question correctly. On the other hand, an item with a high IF index, say .95, would indicate that 95% of the sample responded to the question accurately. It can be said therefore that an item with a small IF index means the question is very difficult and an item with a high IF index indicates that the question is very easy. The ID analysis is useful to indicate the degree to which an item separates high scores from low scores of the test. To do the ID analysis, the samples' result has to be lined up, beginning with their individual item responses and ending with their total scores in descending order. Then two groups will be identified: upper scorers and lower scorers. "The upper and lower groups are sometimes defined as the upper and lower third, or 33%" (Brown, 1996: 67). Brown (op. cit.) adds that the decision as to which way to define these two groups is often a practical matter because groups of samples do not always come in nice neat numbers that are divisible by three. Brown (op. cit.) even finds some instances where 25% was used in calculating ID. There are many ways of calculating an ID index but one of the easiest ways, as suggested by Alderson et.al (1996) and Brown (op. cit.), is by subtracting the IF for the lower group from the IF for the upper group on each item as follows: (taken from Brown, 1996: 67)

ID = IF (upper) - IF (lower)
where ID = item discrimination for an individual item
IF (upper) = item facility for the upper group on the whole test
IF (lower) = item facility for the lower group on the whole test

The job of improving the test particularly for multiple-choice items may not be finished if the DE analysis is not conducted. The selection of the distractors, which are the options that will be counted as incorrect, is very important because they will

divert the examinees from the correct answer if they do not know which is the correct one. The primary goal of DE analysis is to examine the degree to which the distractors are attracting students. To do this, the percentages of each option functioning in the question are calculated and then are analysed.

4.7 Pretesting the preliminary test

Pretesting the preliminary test takes two forms: the first involves a small number of respondents, is done more or less informally with individuals and small groups and involves collecting mostly qualitative feedback; the second is more formal and involves larger groups of respondents from various levels and involves collecting quantitative feedback. The major purpose of the pretesting, as stated in the literature above, is to obtain face and content validity for the test, to find out the most difficult or easiest questions, and to check the time provided for every test and every section of the test and so on.

4.7.1 The pilot experiment of the preliminary test

The preliminary test was tried for the first time as a pilot experiment on a small group of graduate students in the Department of Islamic and Middle Eastern Studies (IMES) at the University of Edinburgh (see Appendix A.2.1: 440-453 for the details of the draft test). The purpose of this pilot experiment was to:

(a) obtain some information about the test's validity in terms of content and face validity, ease of administration, suitable duration for the tests, and ease of scoring although conclusions achieved at this stage are viewed as tentative only,

(b) gain evidence of the differing difficulties of the sub-tests by inter-test correlation. This can be done in this pilot experiment because the respondents who answer the questions as described below are from the higher level of Arabic,

(c) establish the discriminating power of the tests by comparing the results achieved by the samples of this group, and

(d) obtain information about the suitability of the length of pauses between segments in the Dictation Test.

Two groups of students were identified for the sampling purpose. The first group consisted of five postgraduate students in the Department who are native speakers of Arabic. They represent the following countries:

Country:	**N:**
Saudi Arabia	2
Jordan	2
Libya	1
Total	**5**

The selection of these samples was suggested by my supervisor. All except one are studying translation for the Ph.D. degrees. The respondents of this group were briefed about the actual target groups of the test: about their qualifications, their level of Arabic, the Arabic syllabus that they have covered before sitting the intended test and anything else related to the subject matter. To help them get a clearer picture of the test's resources and references, I showed them the Arabic syllabus of the Year One at AIS, which has been discussed in Chapter Two. The second group was smaller, consisted of four Malaysians. They graduated from the AIS specialising in *Sharī`a* and two of them are currently studying *Sharī`a* in the Department of IMES

for Ph.D degrees. The other two are the wives of these two students, one of whom has obtained the Diploma of Education from a university in Malaysia and the second of whom is considering applying for a place for a Masters degree in the Department of IMES. Since they are graduates of the same Academy as the intended candidates, I found that there was no need to brief the respondents of this group about the details of the preliminary test. Furthermore, the syllabus of Arabic during their time at the AIS was almost the same as the current one.

Both groups were briefed personally on how to answer the questions. They were reminded that they had to stick to the time allocated for the whole test and for every part of it. This means that if they could not finish the test or any section of it within the allocated time, they were to make a note in their answer paper. Then they were to continue to answer the rest of the questions in that particular section so that we could analyse the answers to obtain the item difficulties of the remaining questions. For example, in Part A of the Reading Test, the candidates are given ten minutes to answer ten questions including reading the texts of the questions. If, for example, the time is called, and he or she has just finished question seven, he or she should note in the answer paper ten minutes beside the number seven in the answer sheet and then continue to answer the three remaining questions. On the other hand, if they finish before time, they have to indicate at the end of the question number the time they finish answering those questions. The purpose of indicating the time is to determine whether the time suggested for every part of the question and for the whole test is too long or too short. It is assumed that if both groups are unable to answer all the questions within the time limit, the examinees will not be able to do so either.

4.7.1.1 Analysis of face validity of the test

From nine sets of questions distributed to the respondents, I received six. All respondents from the second group returned all test papers and only two from the first group returned the answer papers. However, one respondent from the first group gave a very valuable written commentary on the test items. I then personally met every respondent to hear their comments on the test: the format, the length of the test, the level of the test, etc. The summary of their comments on every sub-test together with the written comments by the respondent mentioned above are as follows:

(i) Reading Test:

(a) The word *khawāṭir* in the question 5 (a) (see Appendix A.2.1: 441) is not appropriate to the context. It is suggested that this word is replaced with the word *makhāṭir* or *muḍar*.

(b) The word *yurabbūhum* in line four in text four (see Appendix A.2.1: 441) should be replaced by the word *yuṭ`imuhum* .

(c) The word *aḍmanū* in question 10 (see Appendix A.2.1: 442) is morphologically wrong and should be changed to *ḍaminū* (without the affix *alif*.)

(d) The text for questions 21 to 25 (see Appendix A.2.1: 443) is not fully in line with the teaching of Islam. It was suggested that this text is replaced by another text that does not culturally contradict the teaching of Islam.

(e) The word *lā* in question 27 and 28 (see Appendix A.2.1: 443-44) is not appropriate to the context and it was suggested that it be replaced by the word *lam*.

(f) The word *biba`īd* which comes after question 54 (see Appendix A.2.1: 444) is not an Arabic style. The correct word is the word *bi`azīz*.

(g) The use of the word *al-afḍal* which comes immediately after question 85 (see Appendix A.2.1: 445) is not appropriate. The respondent argues that *al-afḍal* means 'preferable' whereas in that particular context, it means 'obligatory'. It was suggested that the whole structure of the sentence should be changed.

(h) Another comment made by the respondents concerns the format of the question. According to them, the format of question 9 (see Appendix A.2.1: 441-42), which is multiple choice with more than one correct statement, is not familiar to Arab students.

(i) The first text for the cloze test in the Reading Test is very difficult (see Appendix A.2.1: 444). It was pointed out that new students may not be able to answer these questions.

Except for suggestions (d), (h) and (i), all of these suggestions have been implemented in preparing the revised text for the Reading Test. Suggestion (d) was not implemented because the text as a whole is appropriate for the purpose of testing students' ability in reading and vocabulary. The content of the text is neither against the teaching of Islam nor does it contradict any Islamic principles. Suggestion (h) was not followed because the test is aimed at non-native speakers of Arabic, in this case Malaysians. To the best of my knowledge, Malaysians are familiar with this format and this can be analysed during the pilot surveys in the discussion below. If the pilot surveys show that this particular question is confusing, it will be removed from the final test. For suggestion (i), the decision whether or not to omit the first text of the cloze section will only be made after the analysis of the result has been made.

(ii) Grammar Test

A few comments were made regarding the Grammar Test which can be summarised as follows:

(a) the use of the root word *ḥaḍara* in Question 24 (see Appendix A.2.1: 448) is not appropriate. The correct root word should be *dhahaba.*

(b) the word *ista`artu* in Question 38 (see Appendix A.2.1: 449) should contain the pronoun *hā* which belongs to the word *kutub*. Therefore it was suggested that the word *ista`artuhā* should be used instead of *ista`artu.*

(c) Question 57 (see Appendix A.2.1: 451) which uses the word *`asā* repeats question 30 which asks the same question.

All of these comments and suggestions were implemented in preparing the revised text for the pilot survey.

(iii) Dictation and Essay Tests (see Appendix A.2.1: 452-53)

No comment was made on either test apart from the time limit which will be discussed separately below. The respondents indicated that the content of the test is suitable for the target sample. One of the essays written by one of the respondents will be used as a sample for marking scheme purposes.

4.7.1.2 Analysis on time allocated for the test

As required in the instruction accompanying the test, the respondents provided the following information regarding the time limit.

(i) Reading Test

Part One (see Appendix A.2.1: 440-42): All except one respondent agreed that ten minutes is enough to answer the questions in this part. Most of them even

completed Part One in eight minutes.

Part Two (see Appendix A.2.1: 442-44): Three respondents answered questions in this part in 17 minutes. Another two respondents completed this part in more or less 20 minutes and only one respondent finished it after time. She took 22 minutes to answer questions in this part. However, the respondents indicated that based on their experience when answering the test, if the allocated time remains unchanged in the final test, the examinees may not be able to complete this part. This is because some texts are too long and reading these texts alone takes half of the time allocated for this part.

Part Three (see Appendix A.2.1: 444-45): All except one respondent found that the 20 minutes allocated for this part is not enough. They argued that to answer questions in this part, examinees have to understand first what the texts are about. In other words, they have to read the whole text first before moving to fill in the blank spaces. This task took about half of the time allocated for this part. Some respondents even argued that some questions, eg. nos. 39, 41, 42, 62, 76, have more than one answer, which made them to think deeply of which answer to give. As a result, five of them suggested that if both texts are retained for the final exercise, the time for this part should be no less than 35 minutes. Another two suggested that 30 to 35 minutes are enough for this part.

(ii) The Grammar Test

Part One (see Appendix A.2.1: 446-50): All respondents said that they finished this part when or before the time was called. Three respondents completed this part in 26 minutes, two respondents completed it in 28 minutes and one finished this part just when the time was called. They stated that there should not be a

problem if the suggested time remains unchanged in the final test.

Part Two (see Appendix A.2.1: 450-51): As in Part One, all respondents completed this part within the time limit. They pointed out that the issue was not the time factor but whether the candidates would be able to provide the correct answers for the statements they think are false. They concluded that there is no need to extend the time limit for this part as the matter relates to the content of the test and not to the time itself.

(iii) The Dictation Test (see Appendix A.2.1: 453)

As mentioned above, the purpose of trying out the Dictation Test was to obtain information about the suitability of the length of pauses between segments. Out of six respondents who returned the answer sheets, five conducted the Dictation Tests, while the other made written comments on the item's text. Four respondents conducted the test among themselves; two of them read the text and the other two wrote it. Another respondent conducted it with her sister: she read the text and her sister wrote it. All of them agreed that the length of pauses between segments is suitable. They added that reading the text three times is more than enough. Some of them said that the two minutes pause at the end of the third reading was too long and suggested that one minute would be good enough.

(iv) The Essay Test (see Appendix A.2.1: 452)

All respondents wrote an essay on a given topic. They thought that the 30 minutes was enough because they were provided with some major points to cover. Their tasks were only to arrange these points and to think about the style to be used. In addition, the candidates were required to write the essay in the provided answer papers only, and no maximum or minimum number of words was required. Therefore

candidates do not have to worry about the length of the essay they write. With regard to format, the candidates have a choice whether to write their essay in letter, descriptive, or dialogue format, or any other format they wish, using the main points given in the question paper.

4.7.1.3 Analysis of content validity of the test

The answers were analysed by computer using the Statistical Package for Social Sciences (SPSS). The following are the findings of the analysis:

4.7.1.3.1 Item facility analysis

The analysis below attempts to investigate the item facility (IF) of the Reading and Grammar Tests based on the data in Table 4-1 below.

4.7.1.3.1.1 IF for the Reading Test

Part One: There are ten questions in Part One with multiple choice answer format. Eight questions have an IF index of 1.00. This means all samples answered these eight questions correctly. Question 4 (see Appendix A.2.1: 441) has an IF index of .50 which indicates that only three respondents answered the question correctly. The difficulty in this question is not related to the content of the test. An interview with the respondents reveals that some of them are confused by the distractors *ba'* and *jim*. However, I decided that this question should remain unchanged for the pilot test to see whether or not the samples can answer the question correctly. Question 9 (see Appendix A.2.1: 441-42) has an IF index of .68 because two of the respondents, who are Arab students, are not familiar with the format of the question. This question also remains unchanged because the real test is targeted at non-native speakers of

Arabic who are, to my knowledge, familiar with the format of the question.

Part Two: There are twenty questions in Part Two in a True-False format. Twelve questions have an IF index of 1.00 which means all the samples have successfully chosen the correct answers to the questions. Question 11 (see Appendix A.2.1: 442) has an IF index of .50 only. Interviews with respondents revealed that the wording of the question was confusing. Therefore, the wording has to be changed for the pilot test. Question 28 (see Appendix A.2.1: 444) has an even lower IF index: .33. The respondents claim that the statement was not confusing but their understanding of the earlier text was the reason why they chose the incorrect answer. Therefore the question was considered to have a high content validity and remained unchanged for the pilot test purpose.

Part Three: This part is a cloze test and has two texts: text one covers questions 31 to 54; text two covers questions 55 to 99. The analysis of the IF indices for the first text reveals that the majority of the questions in this part are very difficult. Out of 24 questions (see Appendix A.2.1: 444), 13 have an IF index of .50 and below (three respondents or less answered correctly): 3 questions have an IF index of .00 (none of the respondents supply the correct answer), 5 questions have an IF index of .17 (one respondent answered the question correctly), 4 questions have an IF index of .33 (two respondents answered correctly), and 1 question has an IF index of .50 (three respondents answered correctly).

Text two has 44 questions (see Appendix A.2.1: 444-45). In terms of item difficulty, we may conclude that the questions in this part are easier than the questions in the first text. There are 31 IF indices of above .50, which means that more than three respondents answered the questions correctly: 12 questions have an IF index of

1.00, 11 questions have an IF index of .83 (five answered correctly), and 8 questions have an IF of .67 (four answered correctly). There are only thirteen questions which have an IF of .50 and below: 3 questions have an IF of .50, 9 questions have an IF index of .33, and 1 question has an IF index of .17. None of the questions in this part has an IF of 0.00 (all respondents supplied incorrect answers). Even though one question which has an IF index of .17 (Question n 62, see Appendix A.2.1: 444) could be interpreted as difficult, we cannot remove it from the text because it leads to the second paragraph whereby omitting the sentence of the question will lead to the next paragraph being ambigious.

We establish from the above analysis that the first text is very difficult and therefore is not suitable for the First Year students at the AIS. Four respondents who were formerly students at the AIS agreed that the content of the text was too difficult and therefore was not suitable for the target sample. Even the other two respondents who are native speakers of Arabic also agreed with their colleagues that the text was too difficult for the target samples. They added that apart from the content of the text itself, the deletion rate of the missing word, which in the first text was every fifth word, may contribute to the difficulty of the test question. This was not the case for the second text because the deletion rate here was every sixth word. As a result, I have decided to drop the first text for the pilot survey.

4.7.1.3.1.2 IF for the Grammar Test

There are 65 questions in the Grammar Test. The questions are divided into two parts: Part A consists of 50 questions with multiple choice format and Part B consists of 15 questions in a true-false format. The details of the contents of the test

have been stated earlier in chapter three (see 3.3.2.1). What follows is an analysis of the IF of this test based on the results in Table 4-1 below:

Table 4-1: Item statistics (N=6)

Item no.	Item Facility (IF) (N=6)	Item Facility (IF) (N=6)
	Reading Test	Grammar Test
1	1.00	1.0
2	1.00	1.00
3	1.00	1.00
4	.50	1.00
5	1.00	.83
6	1.00	1.00
7	1.00	1.00
8	1.00	1.00
9	.68	.67
10	1.00	.67
11	.50	.83
12	.68	1.00
13	1.00	.67
14	1.00	.33
15	1.00	.67
16	.33	1.00
17	1.00	.83
18	.83	1.00
19	.83	1.00
20	1.00	.50
21	.68	1.00
22	1.00	.83
23	1.00	1.00
24	1.00	.67
25	.83	1.00
26	1.00	.83
27	1.00	.33
28	.33	.83
29	1.00	.67
30	1.00	.17
31	.83	.17
32	.17	.83
33	1.00	.67
34	.17	.83
35	.33	.83
36	.00	.67
37	.33	.67
38	.83	.83
39	.00	.83
40	1.00	.50
41	.67	.83
42	.67	.83
43	1.00	.83
44	.17	.83
45	.00	.67
46	.50	.83
47	.17	.83
48	1.00	1.00
49	.33	.83
50	1.00	.83
51	.17	1.00
52	.33	.83
53	.50	.67
54	.67	1.00
55	.83	.83
56	.67	1.00
57	.33	.83
58	1.00	.67
59	1.00	.50
60	.83	.83
61	1.00	1.00
62	.17	.83
63	1.00	1.00
64	.33	1.00
65	.67	.33
66	.83	
67	.33	
68	.83	
69	.83	
70	.83	
71	1.00	
72	.67	
73	.83	
74	.33	
75	.33	
76	.50	
77	.83	
78	.83	
79	1.00	
80	1.00	
81	.50	
82	1.00	
83	1.00	
84	1.00	
85	.50	
86	.83	
87	1.00	
88	.67	
89	1.00	
90	.33	
91	.33	
92	.33	
93	1.00	
94	.33	
95	.67	

96	.83	
97	.67	
98	.67	

99	.67	

Part One: As stated earlier in 3.3.2.1 in Chapter Three, there are 36 questions for syntax and 14 items for morphology in Part One (see Appendix A.2.1: 446-50). Out of these 36 items, 6 questions for *mabniy* and *mu`rab* (Indeclined and Declined), *nakira* and *ma`rifa* (Indefinite and Definite), and *mubtada'* and *khabar* (Subject and Predicate); 10 questions for *inna* and its sisters; and 8 questions for *kāna* and its sisters. The analysis reveals that the IF indices for declined and indeclined are as follows: 2 questions have an IF index of 1.00 (question nos. 7 and 23 (see Appendix A.2.1: 447, 448), 3 questions have an IF index of .83 (Questions 34, 35, 42, see Appendix A.2.1: 449, 450), and 1 question has an IF index of .50 (Question 40, see Appendix A.2.1: 449). At this stage, we may say that we can proceed with these questions for the pilot survey. The IF indices for indefinite and definite questions are as follows: 3 questions have an IF index of 1.00 (Questions 1, 2, and 3, see Appendix A.2.1: 446) and the other 3 questions have an IF index of .83 (Questions 22, 38, and 41, see Appendix A.2.1: 448,449). Three questions for Subject and Predicate have an IF index of 1.00 (Questions 3, 4, 16, see Appendix A.2.1: 446, 447) and each question has an IF index of .83, .67, and .50 (Questions 38, 24 and 20, see Appendix A.2.1: 448, 449). It is observed that only half of the respondents answered Question 20 correctly. An interview with the respondents reveals that they were confused by the question. The use of the word *lā* after the pronoun *antumā* makes some of them think that the word *lā* is of the *nāhiya* type (prohibition). Therefore, I changed the pronoun *antumā* with two nouns, *al muslim wa al muslima*, for the pilot survey test

to avoid this confusion.

Part Two: As shown in Table 4-1, Part B (see Appendix A.2.1: 450-51) has 15 questions: 6 questions are related to syntax and 9 questions are related to morphology. The IF analysis reveals that only 2 questions have an IF index of .50 or below: these are Questions 59 and 65 (see Appendix A.2.1: 451) which have the IF indices of .50 and .33 respectively. Both questions seem to be difficult for new students at the AIS. In addition, Question 59 (see Appendix A.2.1: 451) is a false statement which means the candidates have to provide the correct answer for it. Question 65 seems to be very difficult also even though it is a true statement that does not require candidates to provide an alternative answer. Apart from these two questions, I noticed that the time allocated for this part was relatively short because the candidates have to provide the correct answers for statements they think are not true. It is difficult to increase the time limit because the length of time for the whole tests has already been more than two hours. Therefore I find that some other questions in this part need to be removed or at least altered. Question 52 is related to a very low level of Taxonomy category, i.e. knowledge. In addition the statement in that question is too general. Questions 53 and 60 seem to overlap with questions 19 and 31 (see Appendix A.2.1: 448-49) in Part One respectively. Thus I decided to remove five questions, i.e. Questions 52, 53, 59, 60, and 65 for a new version of the test. I have also made alterations to some of the items in this part. The word *`ādā* in Question 64 and the word *`aṣā* in Question 57 (see Appendix A.2.1: 451), which later become Questions 52 and 56 respectively in the new version of the test, have been removed.

It should be noted that in analysing the result of this pilot experiment, I have not simply concluded that an item with a high index of an IF, say .83 or 1.00, is very easy, because the academic level of the respondents, as mentioned above, is very much higher than the target samples in the real test. It could be true, however, as a general conclusion that an item with a low IF index could be assumed to be a very difficult question.

4.8 Fieldwork in Jordan and Malaysia

4.8.1 Fieldwork in Jordan

My supervisor helped to arrange permission for me to conduct a pilot test at the University of Jordan in Amman. I decided to run a pilot test in Jordan for two reasons: to try out the test items on Malaysian students in Jordan, and to try out the same materials on native speakers of Arabic who are also students at the university. With regard to the second reason, Alderson *et. al* (1996) point out that "...suitably defined and selected native speakers is an important aspect of a test on which data ought to be gathered" (p. 97) However, they admit that the issue of pretesting the foreign language tests using native speakers is still controversial and is considered to have some complexities, as discussed by Angoff and Sharon 1971; Alderson 1980; Davies 1991, and Hamilton, Lopes, McNamara and Sheridan 1993. Alderson *et al.* (op. cit.) add that if such a test is not piloted on native speakers, the danger may be that "...test writers may write items which follow the rules of the language, but do not reflect native speaker usage" (p. 97). Being a non-native speaker, I find that it is extremely important for me to try out the test items that I have constructed on native

speakers of Arabic. One very important aspect of piloting test items of foreign language test on native speakers is that since most candidates who are not native speakers cannot be expected to perform to such a high level as educated native speakers, any items -in terms of language usage and not some other factor such as format etc.-, which turn out to be too difficult for native speakers should be omitted.

I arrived in Jordan on 26th March, 1998. In a meeting with Malaysian student representatives in Amman one day after my arrival in the city, I was told that there were students at the University of Jordan in Amman and at the University of *Āl-Al-Bait*, in Mafraq who were willing to be a sample for my pilot study. After another meeting with the Dean of the Language Centre and the Coordinator of the Science of *Hadīth* in the Faculty of *Uṣūluddīn*, both of them at the University of Jordan, I was given permission to conduct a pilot test on a number of students in both Faculties.

(a) Background of the samples

All samples are in Year One or Year Two of their study at university. Samples from Malaysia hold the Malaysian Certificate of Education (M.C.E.) which is equivalent to O Level certificate. Some of them have also completed their two year Arabic language programme at the Language Centre at the University of Jordan before starting their course at the university. The Arab students hold the *tawjīhī* certificate (Secondary School Examination). The number of samples involved in the pilot study differs from one test to another. Therefore the analysis of the pilot test will be covered under the discussion of each respective sub-test.

(b) Sub-tests being tested

Three sub-tests were conducted on the Malaysian samples: Reading, Grammar and Essay Tests. Two sub-tests, reading and grammar, were conducted on the Arab

students. The Dictation Test was not administered in Jordan, due to the facility problem. Regrettably, the materials for the listening test were not obtained in Jordan. It is hoped that materials for this test can be obtained in Malaysia.

The make-up of Arab samples for the Reading Test is 43 (64.2%) out of 67 while the rest are Malaysians. In terms of the subject specialisation, 45 (67.2%) of the samples specialised in Arabic while the other 22 (32.8%) specialised in *Uṣūluddīn*. From the total number of samples from Arab students, 21 are specialised in Arabic.

The population for the Grammar Test in terms of nationality and subject specialisation is as follows: 31 samples (30.3%) are Arab and 46 samples (59.7) are Malaysians which comes to a total of 77 samples. All samples involved in the Grammar Test are specialising in Arabic language. As for the Essay Test, all samples (23) are Malaysian.

4.8.2 Fieldwork in Malaysia

Immediately after completing my fieldwork in Jordan, I left for Malaysia to continue the pilot study and then to administer the final test. I arrived in Kuala Lumpur, the capital of Malaysia on the 1st April, 1998. I had to obtain written permission from the Ministry of Education before I could proceed with my intended research. Therefore I was able to go to schools only in the middle of April, since the approval letter was delayed for two weeks. Unfortunately, the period between the third week of April and the first week of May was allocated for the first term school examination. I contacted two religious boarding schools outside Kuala Lumpur. Both agreed to allow me to conduct the pilot test in their schools after the first week of May.

(a) Background of the samples

The schools that were selected for this study are Sultan Abdul Aziz Shah Islamic College (SAASIC) in Klang and Federal Islamic School of Labu (FISL) in Seremban. All samples were in the Upper Sixth Form and will be sitting for the Higher Certificate of Education (H.C.E.), which is roughly equivalent to A Level, at the end of the year. The total number of the samples from SAAIC was 79 and the total number of samples from FISL was 44, which gave a total of 123 samples. In terms of academic achievement, these samples were among the highest performers in the country. Since all samples from both schools did the same tests and all of them had the same background, I did not separate them into two groups. However, we can still identify the samples from each school because every sample has his or her own identification number (ID).

(b) Sub-test being tested

Three sub-tests were conducted on the samples from SAAIC and FISL: Reading, Grammar and Dictation Tests. The Essay Test was not administered, due to its unsuitability for this group. Regrettably, the materials for the listening test, which suited the content of the syllabus at the AIS, were not available for the purposes of this research. Due to the lack of time available to prepare new materials for the listening test, I decided to drop this sub-test from this research.

4.9 Pilot test administration

4.9.1 Pilot test administration in Jordan

The pilot test for every subject was administered separately. The first test was administered at the *Āl-Al-Bait* University for Reading and Grammar Tests. The test administrator was a Malaysian postgraduate student at the university. He was briefed on how to instruct the students regarding the procedures of the test including the time limit. The second pilot test was administered in the Faculty of *Uṣūluddīn* at the University of Jordan for the Reading Test. The test administrator was the Coordinator of the Science of *Ḥadīth* at the Faculty. The third pilot test was administered in the Faculty of Arts at the same university for Reading and Grammar Tests. The coordinator was the Dean of the Language Centre himself.

4.10 Pilot test administration in Malaysia

I personally administered both pilot tests in Malaysia. The first pilot was administered on Wednesday 6th May. The sample consisted of Upper Sixth form students from FISL. The total number in the sample was 44. The second pilot test was administered on Friday 8th May, i.e. two days after the first pilot administration. 79 samples took the test and all of them were Upper Sixth form students from SAAIC.

With regard to time, students were asked to write down on their answer sheet the time at which they finished, if they finished before the time was up. For example, the time allocated for Part One in the Reading Test is 10 minutes. If they completed this part in 7 minutes, they were to write on their answer sheet beside the last question

in Part One the time at which they finished, i.e. 7 minutes. On the other hand, if they could not finish the test within the time limit, they would be allowed to continue answering the rest of the questions. However they had to indicate on their answer sheet the question number at which they had arrived when the time was up. For example, the time allocated for Part B in the Reading Test was 20 minutes. If the time was called and a student had just completed question 17, he or she was to note beside the question number 17 the time, that is 20 minutes, and then continue with the last three questions. However, extra time allocated to the students for any part was not more than 5 minutes. The purpose of asking the students to write down the time was to get feedback from the samples as to whether or not the time allocated for each part of the test was enough. At the same time, extra time was given on the basis that samples answer all questions so that all items in the test papers could be analysed. (For the details of the second draft of the sub-tests, see Appendix A.2.2: 455-468)

4.11 Findings of the pilot test

4.11.1 Descriptive statistics

In this section, I will display the descriptive statistics resulting solely from the data sample I collected in the pilot test. The purpose of displaying the data in the descriptive statistics is to help us get a clearer overall picture of a data set.

4.11.1.1 Descriptive statistics of samples from Jordan

4.11.1.1.1 The Reading Test

Table 4-2 below shows the descriptive statistics of the samples from Jordan

for the Reading Test:

Table 4-2: Descriptive statistics of samples from Jordan for the Reading Test

N	Valid	67
	Missing	**0**
Mean		**48.40**
Median		**54.00**
Mode		**59**
Std. Deviation		**13.97**
Variance		**195.06**
Range		**49**
Minimum		**16**
Maximum		**65**
Sum		**3243**

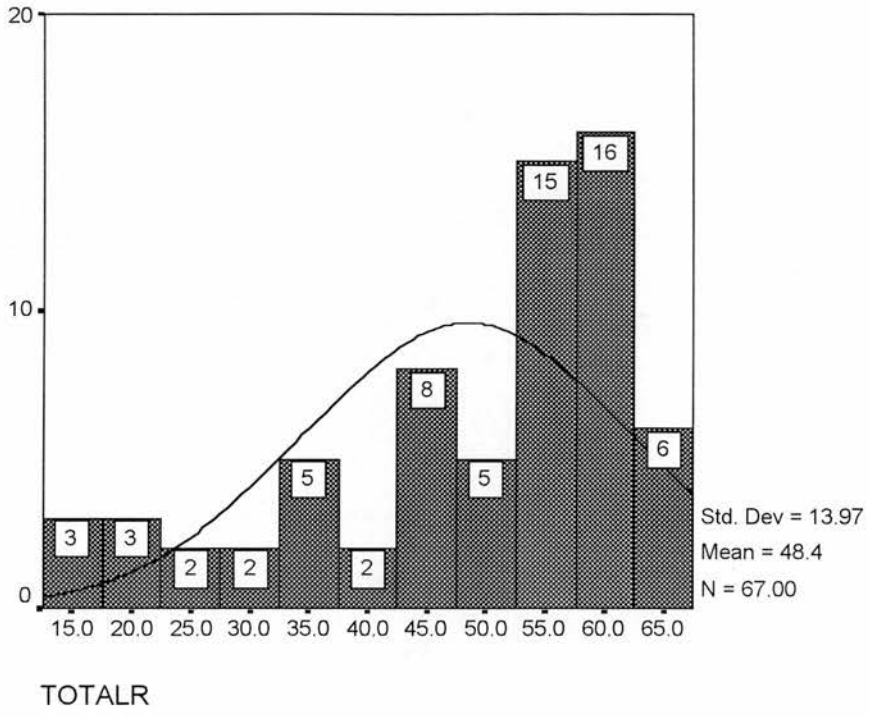
From Table 4-2, we observe the statistical data for central tendency and dispersion. We start the discussion with the former. The mean, 48.40, tells us the average score of the test based on the test population. In this case, SPSS adds up all the individual total scores I compute, i.e. 3243 and divides the total by 67 which is symbolised by N = 67. It should be noted here that the number of items that were on the Reading Test is 75 and is symbolised by $k = 75$. The highest score of the test is also 75 because 1 item is equal to 1 mark. The mean, 48.40, indicates at first glance that this population found the test slightly easy as it is above the 50% mark (precisely 64.53%), and that there are probably more candidates situated towards the top of the distribution than the bottom end. The second statistical data for the central tendency is the median. The median, 54, is that point below which 50% of the scores fall and above which 50% fall. The median, which is slightly above the mean, is another indication that the test is not difficult for the sample. The median also indicates that the majority of the candidates score relatively higher marks. The third is the mode. The mode, 59.00, tells us about the scores that occur most frequently. Even though

the midpoint does not appear in Table 4-2 above, we can calculate the midpoint by adding up the highest and the lowest scores and then dividing them by two. The highest score in the Reading Test as shown in Table 4-2 is 65 and the lowest score is 16. Therefore the midpoint for the test is 40.5.

The second measure in descriptive analysis is dispersion. If the central tendency shows how the scores cluster together, the dispersion shows how widely the scores are spread out. The first statistical measure for dispersion is standard deviation (SD). As shown in Table 4-2 above, the SD for this test is 13.97. From the SD and the mean in this test, I calculate the number of SDs which will fit the distributions that might be described as normal or skewed. 68% of the test population would be found within 13.97 $\pm$ (plus minus) 1 SD, that is 34.43 to 62.37; 95% within 13.97 $\pm$ 2 SDs, that is 20.46 to 76.34; 99.7% within 13.97 $\pm$ 3 SDs that is 6.49 to 90.31. From this calculation, we can fit in 1 SD only on the + (plus) side of the mean which would account for 68% of this population and 2 SDs on the – (minus) side of the mean which could account for 95% of this population. We could therefore describe this distribution as more skewed towards the top than towards the bottom end of the distribution. In non technical terms, a distribution is skewed when the scores are "scrunched up" either towards the higher scores or towards the lower scores. Such distributions characteristically have a tail that points in one of the two possible directions: the lower scores (-) which is termed *negatively skewed* or the higher scores (+) which is termed *positively skewed* (Crocker and Algina, 1986; Anastasi 1988; and Brown, 1988). If the distribution is negatively skewed, it means that most of the candidates scored well. On the other hand, if the distribution is positively skewed, it means that most of the candidates scored poorly. In order to see whether the

distribution in the Reading Test for the samples in Jordan is negatively or positively skewed, I use SPSS to create a graphic representation.

Figure 4-1: Histogram of the Reading Test (Jordan)



The histogram in Figure 4-1 shows that the distribution is negatively skewed. This indicates that more candidates score high marks than lower marks. The line across the histogram is the normal curve line. We can see from Figure 4-1 that the ends of the normal curve line disappear off the histogram not exactly at 0 and 65 but higher up. This confirms the number that fit the distribution I calculated above concerning the 3 SDs on the negative and positive sides, that is 6.49 to 90.31.

The second statistical measure for dispersion is the variance. From Table 4-2 also, we can see the test variance, that is 195.06. This figure shows us the average of the squared differences of students' scores from the mean. The last statistical measure for the dispersion as shown in Table 4-2 is the range. The range is the number of points between the highest score and the lowest score. With the highest score at 65 and the lowest at 16, we can calculate the range as 49.

4.11.1.1.2 The Grammar Test

Table 4-3 below shows the descriptive statistics of the samples from Jordan for the Grammar Test:

Table 4-3: Descriptive statistics for the Grammar Test (Jordan)

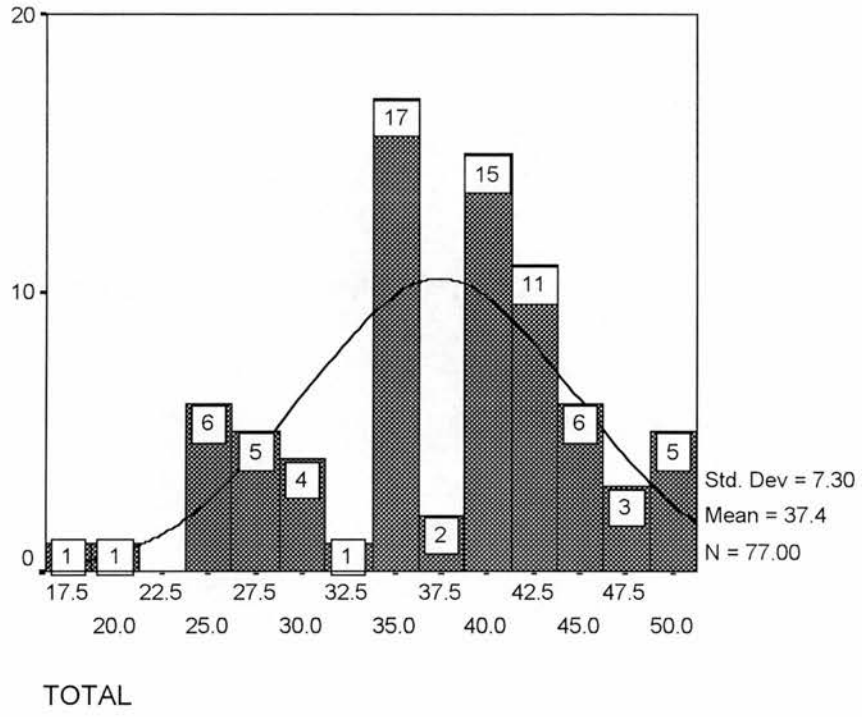
N	Valid	77
	Missing	**0**
Mean		**37.36**
Median		**39.00**
Mode		**36**[a]
Std. Deviation		**7.30**
Variance		**53.31**
Range		**32**
Minimum		**18**
Maximum		**50**
Sum		**2877**

a Multiple modes exist. The smallest value is shown

From Table 4-3, we note the statistical data for central tendency and

dispersion. The number of items that were on the Grammar Test is 60 (k=60) and the highest mark of the test is also 60 because an item equals one mark. The mean, 37.36, indicates that the samples from Jordan found the test fairly easy as it is above the 50% mark (62.26%). The median, 39.00 is another indication that the overall test is not difficult for the samples. The same indication to show that the test is not difficult is shown by the mode, 36, which is one point above the middle point of the total marks. The last statistical measure for the central tendency is the midpoint. The midpoint of the test is 34. From the statistics, we may conclude that there are probably more candidates situated towards the top of the distribution than the bottom end. With reference to the dispersion, the standard deviation (SD) for this test is 7.30. From the SD and the mean, I calculate the number of SDs which will fit the distributions that might be described as normal or skewed. 68% of the test population would be found within 7.30 $\pm$ 1 SD, that is 30.06 (37.36 – 7.30) to 44.66 (37.36 + 7.30); 95% within 7.30 $\pm$.2 SDs, that is 22.76 to 51.96; and 99.7% within 7.30 $\pm$ 3 SDs, that is 15.46 to 59.26. From this calculation, I can fit 2 SDs on (+) side of the mean which would account for 95% of the population and 3 SDs on the (-)side of the population. I could therefore describe this distribution as skewed towards the top than towards the bottom end of the distribution. The histogram in Figure 4-2 below displays the graphic representation of the distribution in the Grammar Test for samples from Jordan:

Figure 4-2: Histogram of the Grammar Test (Jordan)



The histogram in Figure 4-2 indicates that the distribution is negatively skewed. We can see the normal curve line across the histogram disappear off the histogram not exactly at 50 but higher up on the (+) side but at 0 on the (-) side. This confirms the calculation I made earlier concerning the 3 SDs on both sides (±), that is 15.46 to 59.26. From Table 4-3 also, we can see test variance and the range. The range, 32, provides us some idea of how individuals vary from the central tendency. The test variance, 53.31, gives us some information about the average of the squared differences of candidates' scores from the mean.

4.11.2 Item analysis of the pilot test

Three item analysis instruments will be used in examining the test items: item facility analysis, item discrimination index analysis, and distractor efficiency analysis. As discussed earlier, the facility analysis measures the level of difficulty of an item. On the other hand, the discrimination index analysis measures the extent to which the results of an individual item correlate with results from the whole test (Brown 1996, Alderson *et al.* 1996). These two analyses are useful in order to understand and to improve the effectiveness of item format and contents. Brown (op. cit.) warns however that we should be careful in using these statistical techniques, saying that "...the statistics are only for improving actual test items and are not an end in themselves" (p.64). The distractor efficiency analysis examines the degree to which the distractors are attracting students. The discussion for the above topic will start with the samples from Jordan first, followed by the samples from Malaysia.

4.11.2.1 Item analyses of the pilot test for samples from Jordan

(i) The Reading Test

4.11.2.1.1 Item facility (IF) analysis

Every item in the Reading Test was computed on the samples (N=67). Table 4-4 below shows the IF obtained from the summary of the samples' total score of the Reading Test.

Table 4-4: Item facility for the Reading Test (Jordan)

Item no.	Item Facility (IF)
1	.81
2	.81
3	.67
4	.39
5	.93
6	.91
7	.93
8	.78
9	.55
10	.97
11	.73
12	.70
13	.90
14	.63
15	.86
16	.88
17	.91
18	.88
19	.72
20	.93
21	.64
22	.69
23	.76
24	.82
25	.61
26	.93
27	.54
28	.46
29	.84
30	.81
31	.62
32	.46
33	.34
34	.43
35	.85
36	.80
37	.79
38	.34
39	.73
40	.57
41	.49
42	.67
43	.12
44	.73
45	.66
46	.75
47	.79
48	.62
49	.60
50	.24
51	.12
52	.19
53	.46
54	.54
55	.13
56	.55
57	.51
58	.75
59	.75
60	.81
61	.43
62	.69
63	.79
64	.75
65	.79
66	.49
67	.27
68	.37
69	.78
70	.06
71	.22
72	.67
73	.73
74	.67
75	.67

To analyse the IF, I use the interpretation suggested by Brown (1996). Brown suggests that "...an item with an IF of .27 would be a very difficult question because many more students missed it than answered it correctly" (p.65). Therefore, items are regarded as unsatisfactory if they have low IF indices on the sample (N=67). In Part One (Questions 1-10) (see Appendix A.2.2: 456-57), all IF indices are above .27 which means the items are not difficult. The same happens in Part Two) (Questions 21-30 (see Appendix A.2.2: 457-59) where no item has an IF below .27. The lowest IF index is .46 (Question 28). In Part Three (Questions 31-75) (see Appendix A.2.2: 459-60), several questions have an IF index below .27 (Nos. 50, 51, 52, 55, 70, 71). Item 70 has the lowest IF, i.e. .06 which means only 4 samples answered this question correctly. The pilot study shows that some of the questions have low IF indices because they have more than one possible answer. Selecting one correct answer only to every question when marking, was the main reason for the low IF index of those items. Another factor to mention in analysing IF for items in Part Three is that I cannot omit or drop some questions from the text because the questions are related to each other. Omitting an item will affect the meaning of the whole text or at least particular sentences and paragraphs. Some alteration to the wording or vocabulary may be useful to overcome this problem. However, the final decision on the questions with low IF indices will be taken only after the second pilot test has been conducted on samples from Malaysia.

Apart from the IF analysis, the indices also indicate other interesting points. With regard to Part One (see Appendix A.2.2: 456-57), it may be true to say that there is no difference between the general Arabic texts and religious texts in terms of difficulty or even the use of vocabulary for both types of text. In Part Two (see

Appendix A.2.2: 457-59), the first and the last texts were taken from religious text books for Year One at the AIS while the other three were selected from general Arabic books. From the IF indices above, we may conclude that there is no difference in the samples' score between these two types of text. This may be used as evidence to reject the claim that texts related to religious context are not suitable for a language test.

Another point that is interesting to note here is the format of the question. As discussed earlier, Arab samples are not familiar with the format of Question 9 in Part One (see Appendix A.2.2: 457). Most of them answered it incorrectly. As a result, this item has a relatively low IF index (.55).

4.11.2.1.2 Item discrimination (ID) analysis of the Reading Test

Since the number of samples involved in the pilot test is small, I will use approximately 25% (15 samples) of the total number of samples for the lower and upper groups in calculating the ID for the Reading Test. From the summary of item statistics, 25% of the upper group includes samples with a total mark ranging between 65 and 59. However, only 2 samples from those who obtained 59 marks are included in the upper group to make the number of sample up to 15. In the case of the lower group, 25% of it includes samples with a total mark ranging from 16 to 37. From the summary of the frequency of the total score for the Reading Test, the calculation to divide two groups, upper and lower, was conducted. Then the ID index was calculated for each item as shown in Table 4-5 below:

Table 4-5: Item discrimination index for the Reading Test (Jordan)

Item	IF (upper)	IF (lower)	ID
1	.87	.60	.27
2	.93	.47	.46
3	.87	.20	.67
4	.47	.27	.20
5	.93	.80	.13
6	.93	.80	.13
7	.93	.93	.00
8	.93	.73	.20
9	.60	.53	.07
10	1.00	1.00	.00
11	.73	.60	.13
12	.93	.73	.20
13	1.00	.73	.27
14	.67	.53	.14
15	1.00	.53	.47
16	.93	.80	.13
17	1.00	.93	.07
18	.93	.73	.20
19	.93	.53	.40
20	1.00	1.00	.00
21	.93	.20	.73
22	.87	.33	.54
23	1.00	.33	.67
24	.87	.87	.00
25	1.00	.27	.73
26	1.00	.87	.13
27	.60	.60	.00
28	.40	.33	.07
29	1.00	.73	.27
30	.93	.53	.40
31	1.00	.13	.87
32	.80	.00	.80
33	.67	.34	.33
34	1.00	.80	.20
35	1.00	.47	.53
36	1.00	.47	.53
37	.93	.47	.46
38	.33	.07	.26
39	1.00	.27	.73
40	.87	.33	.54
41	.93	.13	.80
42	.80	.33	.47
43	1.00	.33	.47
44	.93	.13	.80
45	1.00	.07	.93
46	.93	.27	.66
47	1.00	.27	.73
48	1.00	.13	.87
49	1.00	.20	.80
50	.47	.07	.40
51	.20	.00	.20
52	.40	.13	.27
53	.80	.00	.80
54	.73	.00	.73
55	.33	.00	.33
56	.87	.13	.74
57	.67	.07	.60
58	1.00	.20	.80
59	1.00	.20	.80
60	1.00	.33	.67
61	.67	.13	.54
62	1.00	.27	.73
63	1.00	.27	.73
64	1.00	.20	.80
65	1.00	.27	.73
66	.80	.00	.80
67	.40	.00	.40
68	.60	.00	.60
69	1.00	.20	.80
70	.13	.13	.13
71	.40	.07	.33
72	1.00	.13	.87
73	1.00	.13	.87
74	.87	.13	.74
75	1.00	.07	.93

To draw conclusions about the items based on the item discrimination analysis above, Ebel (1979: 267) suggests the following guidelines:

.40 and above	Very good items
.30 to .39	Reasonably good but possibly subject to improvement
.20 to 29	Marginal items, usually needing and being subject to improvement
Below .19	Poor items, to be rejected or improved by revision

Alderson *et al.* (1996) and Brown (1996), however, argue that such a guideline should not be used as a hard and fast rule but rather as an aid in making decisions about which items to reject and which to keep. According to Alderson *et al.* (op. cit.), item writers are often content with ID of .40 or above, but there are no rules as to which IDs are acceptable since "...the possibility of getting high DIs (ID) varies according to the test type and range of ability of the examinees" (p.82). Another factor that should be considered before making a decision on items with a low ID index is whether the items are too easy or too difficult as can be observed from the IF of both groups, i.e. upper and lower groups. From a humanitarian point of view, it is good to keep this kind of item, "...just so the students can get off to a good start" (Brown, op. cit: 71). However, an item with a negative ID shows that something has gone very wrong with such an item and it should be revised or discarded (Noll, *et al.*1979; Ebel, 1979; Alderson, *et al.*1996; Brown, 1996). There is good reason to doubt the value of the contribution made to a test result by items that have negative ID indices. The negative index shows clearly that samples from the lower group are better than samples from the upper group for such items.

With reference to Table 4-5, 5 questions in Part One (see Appendix A.2.2: 456-57) of the Reading Test have ID indices below .20. This indicates that those questions are not discriminating between the upper and the lower groups. However,

the IF indices in Table 4-4 indicate that these items are quite easy. Therefore I intend to retain these questions as a good start for candidates.

In Part Two (see Appendix A.2.2: 457-59), 9 questions have ID indices below .20. IF indices of these questions as shown in Table 4-9 indicate that all of these questions except Question28 have an IF index of above .50. As Alderson *et al.* noted earlier, the range of ability of the samples may contribute to this result. However, the wording of the question and the way the question is phrased may also be a reason for low ID indices. Therefore, I need to revise and make necessary changes to these questions for the future version of the test.

ID indices in Part Three (see Appendix A.2.2: 459-60) clearly discriminate between these two groups. Only one of the ID indices of the questions falls below .19 and 36 questions from the total of 45 are above .40 ranging from .40 to .93.

4.11.2.1.3 Distractor efficiency (DE) analysis

Table 4-6 below summarises the percentage of options made by the samples (N=67) for Part One of the Reading Test.

Table 4-6: Distractor efficiency statistics (N=67)

Items	Options			
	a	b	c	d
1	8	6	5	81*
2	6	81*	3	9
3	19	67*	6	3
4	28	24	38*	9
5	93*	3	3	0
6	8	0	0	91*
7	93*	5	3	0
8	78*	2	5	16
9	24	19	2	55*
10	0	2	2	97*

*correct option

If we consider the DE analysis result in Table 4-6 above, we notice that the table also provides the same item facility indices that were discussed and shown in Table 4-4 earlier. For example, Question 1 (see Appendix A.2.2: 456) which indicates *d* as the correct answer (indicated by an asterisk) has the same IF index as Question 1 in Table 4-4. This analysis, however, provides additional information about the proportion of students who choose each of the options. For instance, in Question 1, nearly 81% or 54 samples chose option *d*. Only 19% went to other options, i.e. *a*, *b*, and *c*. From this result, we can say the degree to which the distractors are attracting the candidates is low. However, it may be the case that the problem does not necessarily occur because of the distractors alone. It could be caused by the question itself: it is so easy sometimes for candidates to pick up the correct answer.

Valuable insights can also be provided by DE analysis when the majority of candidates choose an option regardless of whether the answer is right or wrong. For example, in Question 4 (see Appendix A.2.2: 456), option *c* is the correct answer as indicated by the asterisk, with about 38% or 26 samples choosing it. Oddly, about 52% selected wrong answers, which are option *a* and *b*. In a situation like this, it is important for the researcher to check the original item and examine it carefully in terms of both format and content to ensure there is no element that may divert the candidates from choosing the correct option. By doing this, it is hoped that the IF and ID indices will be increased in future.

In Questions 5, 6, 7, and 10 (see Appendix A.2.2: 456-57), options other than the correct answer do not seem to be very attractive. Question 6 provides an example of an item with two distractors and Questions 5, 7, and 10 provide an item with one

distractor which does not attract any of the candidates in this pilot study. In other words, these distractors are of little value in the process of distracting the candidates. I may decide to revise these options so that they will be more attractive. Alternatively, I will leave the items alone and continue to use the options according to the theory that tampering with an item that is working is foolhardy (Brown, 1996). Questions 3 and 9 (see Appendix A.2.2: 456, 457) look like good questions with well-centred options except for option c in Question 9.

(ii) The Grammar Test

4.11.2.1.4 Item facility (IF) for the Grammar Test

Every item in the Grammar Test was computed on the samples (N=77). Table 4-7 below shows the IF for the Grammar Test obtained from the summary of the samples' answer.

Table 4-7: Item facility for the Grammar Test (N=77)

Item no.	Item Facility (IF)
1	.84
2	.88
3	.91
4	.79
5	.38
6	.64
7	.86
8	.75
9	.53
10	.51
11	.68
12	.82
13	.26
14	.48
15	.48
16	.91
17	.91
18	.90
19	.52
20	.42
21	.75
22	.78
23	.88
24	.87
25	.44
26	.94
27	.31
28	.81
29	.42
30	.42
31	.47
32	.95
33	.44
34	.77
35	.29
36	.62
37	.84
38	.69
39	.78
40	.23
41	.87
42	.62
43	.40
44	.51
45	.13
46	.39
47	.26
48	.87
49	.57
50	.71
51	.91
52	.84
53	.42
54	.26
55	.36
56	.87
57	.10
58	.82
59	.81
60	.35

To analyse the IF for this test, I use the same interpretation as suggested by Brown (1996) when analysing the Reading Test. In Part One (Questions 1-50) (see Appendix A.2.2: 462-66), most IF indices are above .27 which mean the questions are not difficult. 17 questions have IF indices between .80 and .95 which means between 61 and 73 samples answered the items correctly. Only 4 questions have IF indices below .27. These questions are 14 and 47 (.26), 40 (.23), and 45 (.13) (see Appendix A.2.2: 463, 465). Investigation shows that questions which have low IF indices are related to word roots (Questions 14, 47) and declension (Questions 40, 45). Since those questions which have low IF indices are related to the syllabus, I do not intend at this stage to drop them. Morover, the final decision on the items with low IF indices will be taken only after the second pilot test has been analysed on samples from Malaysia.

The IF indices in Part B (see Appendix A.2.2: 466) show that two questions have low IF indices (items 54, 57). Further investigation reveals that both questions have low IF indices because the samples did not provide the correct answer for both of them, even though they knew that the statements in the questions were not true. For example, for Question 54, 22 samples (28.6%) provided half-correct answers. This means that they know the statement in the question is false. However, when giving the answer to this question, they put it wrongly or they were simply unable to provide the correct answer. The same thing happens to Question 57: the samples provided half-correct answers to the question, which results in low IF index for this question.

4.11.2.1.5 Item discrimination (ID) analysis (Jordan)

Since the number of samples involved in the pilot test is relatively small, I will use approximately 25% (19 samples) of the total number of samples (N=77) for both groups, lower and upper, in calculating the ID for the Grammar Test. In the case of the upper group, 25% of the total samples includes samples with a total mark ranging from 42 to 50. However, only 3 samples from those who obtained 42 marks are included in the upper group to make the total number of sample 19. In the case of the lower group, 25% of total samples includes samples with a total mark ranging from 18 to 34. However only one sample from those who obtained 34 marks is included in the lower group to make the total number of samples 19. Table 4-8 below displays the ID for the Grammar Test for samples from Jordan.

Table 4-8: Item discrimination (ID) statistics for the Grammar Test (Jordan)

Item	IF (upper)	IF (lower)	ID
1	.95	.69	.26
2	1.00	.84	.16
3	1.00	.79	.21
4	1.00	.63	.37
5	.74	.00	.74
6	1.00	.16	.84
7	.84	.79	.05
8	.63	.73	-.10
9	.79	.26	.53
10	.79	.42	.37
11	.84	.58	.26
12	.95	.63	.32
13	.95	.84	.11
14	.47	.05	.42
15	.47	.32	.15
16	.95	.79	.16
17	1.00	.79	.21
18	.95	.90	.05
19	.68	.16	.52
20	.63	.32	.31
21	1.00	.47	.53
22	.90	.68	.22
23	1.00	.63	.37
24	.95	.74	.21
25	.84	.16	.68
26	.95	.79	.16
27	.68	.00	.68
28	.95	.58	.37
29	.68	.26	.42
30	.53	.05	.48
31	.68	.11	.57
32	.1.00	.90	.10
33	.68	.32	.36
34	.95	.37	.58
35	.42	.21	.21
36	1.00	.21	.79
37	1.00	.58	.42
38	.95	.32	.63
39	.95	.58	.37
40	.42	.00	.40
41	.95	.90	.05
42	.68	.63	.05
43	.47	.41	.26
44	.79	.11	.68
45	.11	.16	-.05
46	.47	.26	.21
47	.53	.11	.42
48	1.00	.63	.37
49	.84	.26	.58
50	.95	.42	.53
51	.84	.84	.00
52	1.00	.63	.37
53	.53	.32	.21
54	.26	.16	.10
55	.26	.42	-.16
56	.95	.84	.11
57	.32	.05	.27
58	.95	.58	.37
59	.84	.68	.16
60	.31	.32	-.01

To analyse the data in Table 4-8, the guidelines suggested by Ebel (1979) above will be used. To make the analysis of the ID index easier, I summarise the data in Table 4-8 above as follows:

Table 4-9: Summary of the ID indices for the Grammar Test (Jordan)

ID Range	Frequency	Percent
Below .00	5	8.3
.01 - .19	13	21.7
.20 - .29	11	18.3
.30 - .39	11	18.3
.40 above	20	33.3
Total	60	100

ID indices from .20 to 1.00 are considered acceptable in this pilot study. What remains to be discussed is items with an ID of .19 and below. As shown in Table 4-9, 18 questions have an ID below .20. The discussion starts with questions with an ID below .00. These are Questions 8, 45, 51, 55, and 60 (see Appendix A.2.2: 4462, 465, 466). A close investigation of these questions reveals that they are related to the declension in the syllabus for Year One at the AIS. All except one question have item facility (IF) indices of above .35: they are therefore considered not to be difficult (see Table 4-7 for the details of IF of these items). Hence I intend to retain these questions for at least the second tryout in Malaysia before making any decision whether or not to drop these questions in the final version. With regard to Question 45 which has an IF index of 13 and an ID index of -.05, it seems to be the most difficult question under the declension topic. I also intend to retain this question for the second tryout to see whether or not it works with the samples in Malaysia. With reference to item 54, even though the ID index was very low, the IF index was acceptable. As for fifteen questions that have ID indices ranging from .01 to .19, I find that almost all questions have IF indices of above .60. Only one question, i.e. Question 54, has an IF index of .26, one point below the acceptable IF index which

has been suggested by Brown (1996). Since the test is intended to detect low and high ability students so that they can be placed in suitable groups according to their ability and not merely to rank them, I find these questions may be retained for the next stage of the pilot test.

4.11.2.1.6 Distractor efficiency (DE) analysis (Jordan)

As stated earlier, the task of improving the test particularly with regard to multiple-choice items may not be completed if the DE analysis is not conducted. In this section, I will examine the nonfunctioning and malfunctioning distractors of every option in Part One of the Grammar Test with a view to either discarding or revising them for future versions of the test. Table 4-10 below displays the percentage of options made by the samples (N=77) for Part One of the test.

Table 4-10: Distractor efficiency statistics for the Grammar Test (Jordan)

Items	Options			
	a	**b**	**c**	**d**
1	0	12	4	84*
2	0	8	88*	4
3	5	91*	0	1
4	79*	20	1	0
5	12	38*	26	22
6	29	0	5	64*
7	1	5	86*	7
8	0	10	14	75*
9	20	53*	3	21
10	51*	43	4	1
11	22	68*	1	5
12	12	3	82*	4
13	6.5	9	82*	3
14	26*	4	34	34
15	48*	46	3	4
16	7	1	1	91*
17	1	91*	7	1
18	3	7	0	90*
19	1	25	52*	18
20	35	42*	13	10
21	9	8	75*	8
22	17	1	4	78*
23	5	88*	3	4
24	5	3	5	87*
25	44*	4	31	21
26	94*	3	1	0
27	14	20	31*	35
28	81*	7	9	3
29	33	8	18	42*
30	17	42*	1	39
31	47*	31	12	9
32	4	1	0	95*
33	0	20	44	33
34	7	77*	16	0
35	43	12	29*	14
36	17	62*	3	17
37	12	1	1	84*
38	0	25	69*	4
39	78*	8	5	5
40	53	23*	13	5
41	4	4	87*	3
42	9	62*	22	4
43	17	29	40*	10
44	5	23	17	51*
45	13*	17	57	8
46	38	34*	18	7
47	26*	46	25	0
48	9	0	0	87*
49	9	57*	4	26
50	0	1	22	71*

*correct option

From Table 4-10, we observe the proportion of students who choose each of the options. Many valuable insights can be obtained from Table 4-10. Firstly, the above table shows the options chosen by a majority of samples even though they were in fact wrong. Out of 50, 7 questions, 14, 27, 35,40, 45, 46, and 47 (see Appendix A.2.2: 463, 464, 465), have the majority of candidates selecting wrong answers. As stressed earlier in the discussion of DE analysis for the Reading Test, these questions need to be examined in terms of both format and content to identify the element that may divert the candidates from choosing the correct options. With reference to Question 14, the majority of samples chose options *c* and *d*. A close look at Question 14 suggests that this happened because they could not differentiate between the words *ijtāza* (option *a*), **ajāza** (option *c*), and *tajāwaza* (option *d*). Although it was clear that option *a* was the correct answer, option *d* could be the correct option too if the samples considered the word *Islām* to be above the words *al-ḥaḍārat al-rāqiya*. If this question is retained for the final version, the word *tajāwaza* in option *d* should be replaced by another word to avoid confusion. Question 27, which refers to the use of *al-asmā' al-khamsa* in the forms of *fatḥa* and *iḍāfa*, sees the majority of candidates choosing option *d* instead of option *c*. This shows the candidates' ignorance regarding the use of this noun. The same happened to Questions 35 and 40 where the samples did not know the predicate of the subject (item 35) and the predicate of the noun *inna* (item 40). Therefore no modification needs to be made to this item. With reference to Questions 45 and 46, both questions were related to the noun of *kāna*. In Question 45, the majority of candidates (57%) chose option *c* even though the short vowel of *fatḥa* was clearly indicated on the word *jawāb*. It may be assumed at this stage that the candidates were not familiar with the use of *al-ḍamma al-*

muqaddara of the noun of *kāna*. The same happens to Question 46 where only 34% of the candidates chose the correct option. They were unable to find the noun of *kāna* because it appeared after the preposition of *min*. We may conclude at this stage that this item was difficult for candidates at this level. With regard to item 47, the majority of candidates were unable to allocate the root of *ishtāqa*. As a result, most of them (46%) chose option *b* instead of *a*.

Secondly, Table 4-10 also shows that the degree to which some distractors were attracting the candidates is very low. Some distractors did not attract the candidates at all while others had a very low percentage ranging from 1.3 to 13% which means less than 10 candidates chose them. In a normal situation, these distractors need to be revised to make them more attractive for future versions of the test. However, I have to stress here that some distractors cannot be easily replaced by others. For example, Question 6 (see Appendix A.2.2: 462) has distractor *b*, *allatī*, which did not attract any candidate because it was the only option in the feminine form while the other three were in the masculine form. However, it is difficult to find another word that can attract candidates more. We may replace the word *allatī* with *alladhāni* but the possibility of it becoming attractive is still low because the same word, *alladhayni,* which has been used as another distractor in option *c* in the accusative or genitive form, attracted only 4 candidates (5.2%). Another example is Question 2 (see Appendix A.2.2: 462): no candidate chose option *a* and only 3.9% (3 candidates) chose option *d*. However, I think that it will prove too difficult to replace the words *rajulan* (option *a*) and *rijāl* (option *d*) with any other words that are similar to the correct option, i.e. *rajul* (option *c*). Taking the above discussion into consideration, I decided to revise seven only. Table 4-11 below summarises the

options which will be considered for revision:

Table 4-11: Distractors with low percentage for the Grammar Test (Jordan)

item nos.	*Options*
1	a
3	c
4	c, d
9	c
17	d
25	b

Thirdly, Table 4-10 also reveals that some correct options, ranging from 85 to 94%, were able to attract the majority of candidates. This can be seen in Questions 2, 7, 18, 23, 41, and 48 (between 85 to 89%) and Questions 3, 16, 17, 26, and 32 (between 90 to 94%) (see Appendix A.2.2: 462-65). This indicates either that the question was too easy and therefore candidates chose the correct option easily, or that the distractors were not functioning efficiently. However, the final decision as to whether or not to revise these options will be taken only after the analysis of distractor efficiency on the samples from Malaysia has been made.

(iii) The Essay Test

As mentioned earlier, all samples involved in the Essay Test in Jordan were Malaysians. Although the item analyses used for the Reading and Grammar Tests which were discussed earlier are inappropriate for a writing test, this test still needs to be analysed to investigate the following:

(a) whether the given topic extracts the intended sample of language;

(b) whether the given topic is suited to the candidates' language ability and level;

(c) whether the marking scheme that has been drafted during the item writing stage in Chapter Three is usable; and

(d) whether the examiners are able to mark the essays consistently.

It is difficult at this stage to try the test out on a large number of samples because of the time needed to mark the scripts. However, efforts have been made to make sure that the samples (N=23) in Jordan represent a wide range of backgrounds and language levels to ensure that the sample of language produced contains most of the features that will be found in the examinations themselves. To achieve this, I asked the Malaysian students' representative to choose as many samples as possible for the Essay Test from various levels based on their results in the previous examinations.

The pilot test for an essay was administered only after I had left Jordan for Malaysia (see Appendix A.2.2: 467). I asked the Malaysians student representative to send me the answer scripts. After receiving the answer scripts, I quickly read the scripts (k=23) to familiarise myself with the types of writing that the candidates had produced and the problems they had in performing the task. Using the rating scale which was set up in chapter three, I extracted the scripts which represented 'adequate' and 'inadequate' performance. In addition, other problems that are rarely described in rating scales such as bad handwriting, excessively short or long responses, responses which indicate that the candidate misunderstood the task, etc, were analysed at this stage. With regard to the number of scripts, Alderson *et al* (1996) suggest that at least 20 scripts representing various levels of performance should be used. Thus I decided to read all the scripts (k=23). As was described earlier (see Chapter Three: the scoring method), the total possible mark for the essay

was 25: 10 marks were allocated for content and organisation; and 5 marks each were allocated for the use of vocabulary, accuracy in grammar, and the mechanics, which included punctuation, spelling, etc. Later I took the scripts to my colleagues who served as my *standardising committee* to try out the rating scale on these scripts and to see whether the rating scales were usable or not. The standardising committee consists of myself and another two lecturers from the Faculty of Language and Linguistics at the University of Malaya. Both of them teach Arabic language at the AIS. They were given the scripts one after the other and were asked to mark them. They were asked to give the marks for every script based on the rating scales on a separate paper in order to avoid influencing each others' marks. Table 4-12 provides a summary of the descriptive statistics for each rater:

Table 4-12: The summary of descriptive statistics for the Essay Test (Jordan)

nos		ct/or 1	voc 1	grm1	mch1	tR 1	ct/or 2	voc 2	grm2	mch2	tR 2	ct/or 3	voc 3	grm3	mch3	tR 3
1		5	3	3	3	14	6	3	3	4	16	5	3	3	3	14
2		4	2	2	3	11	4	2	2	2	10	4	2	2	2	10
3		7	3	4	4	18	9	4	4	5	22	9	4	4	5	22
4		5	3	3	3	14	7	4	4	5	20	6	3	4	4	17
5		4	2	2	2	10	5	2	3	3	13	5	3	3	3	14
6		3	2	2	2	9	3	1	2	2	8	3	2	2	2	9
7		5	3	3	3	14	7	3	4	4	18	7	4	4	5	20
8		4	2	3	3	12	6	3	4	4	17	6	4	3	4	17
9		4	2	2	2	10	6	3	3	3	15	6	3	3	3	15
10		5	3	3	3	14	7	3	4	4	18	6	3	3	4	16
11		4	2	2	3	11	4	2	2	3	11	4	2	2	2	10
12		6	3	3	4	16	7	4	4	5	20	7	4	4	5	20
13		4	2	2	3	11	3	1	1	2	7	4	2	2	2	10
14		4	2	2	3	11	5	2	3	3	13	4	2	2	2	10
15		5	2	2	3	12	4	2	2	2	10	5	3	2	3	13
16		4	2	2	2	10	4	2	3	3	12	5	3	3	3	14
17		5	3	3	3	14	3	2	2	3	10	3	2	2	3	10
18		7	4	4	5	20	8	4	4	5	21	8	5	4	5	22
19		5	3	3	4	15	7	3	4	5	19	7	4	4	5	20
20		5	3	2	3	13	5	3	3	3	14	5	3	3	3	14
21		4	2	2	2	10	2	2	2	2	8	4	2	2	3	11
22		5	2	3	3	13	3	2	2	3	11	4	2	2	3	11
23		3	1	2	2	8	2	1	2	2	7	3	2	2	2	9
Total	N	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23

Note:

nos. = the number of students
ct/or 1, 2, 3= content/organisation for raters 1,2, and 3
voc 1,2,3 = vocabulary for raters 1, 2, and 3
grm 1, 2, 3 = grammar for raters 1, 2, and 3
mch 1,2,3 = mechanics for raters 1, 2, and 3
tR 1, 2, 3 = total marks for raters 1, 2, and 3

Table 4-12 shows that there were no big differences between the three raters in giving their marks to the samples' work. With reference to the first aspect of writing, i.e. content and organisation (ct/or), we observe that the difference between the marks of the three raters ranged from 1 to 2 marks only. The correlation coefficient (r) between these three raters was relatively high: the r rater 1 and rater 2 was .732; the r between rater 1 and rater 3 was .784; and the r between rater 2 and 3 was .940, all of which were close to 1.00. With reference to the second aspect, i.e. vocabulary, the above data show that the raters differed slightly for sample 8. Rater 1 gave 2 marks, rater 2 gave 3 marks, and rater 3 gave 4 marks. Other than this, if they differed, it was within 1 mark only. The correlation coefficient between the three raters was also high: .676; .781; and .802. The third aspect, i.e. grammar, also indicates that the raters did not differ from each other by more than 1 mark. This resulted in a high correlation between the three raters: the r between rater 1 and rater 2 was .660; the r between rater 1 and 3 was .680; and lastly the r between rater 2 and 3 was .899. The last aspect of writing that was assessed was the mechanics. The r between rater 1 and rater 2 was .711, between rater 1 and rater 3 was .659, and between rater 2 and rater 3 was .871. This high correlation indicates clearly that the raters did not differ among themselves when assessing the essays using the rating scales mentioned earlier. From the data in the above tables and the correlation of the aspects of writing that has been computed, we may assume the correlation among the three raters for the total marks should be high. Table 4-13 below summarises the correlation coefficient of the three raters for total marks:

Table 4-13: Correlation coefficient of total marks of the essay for the three raters

		TTLR1	TTLR2
TTLR2	Pearson Correlation	.801**	1.000
	Sig. (2-tailed)	.000	.
	N	23	23
TTLR3	Pearson Correlation	.792**	.930**
	Sig. (2-tailed)	.000	.000
	N	23	23

**Correlation is significant at the 0.01 level (2-tailed).

From Table 4-13, we observe that the correlation coefficient between the three raters for the total marks of the Essay Test was high: the *r* between rater 1 and rater 2 was .80, the *r* between rater 1 and rater 3 was .79; and lastly the *r* between rater 2 and rater 3 was .93. However, it is interesting to note here that among the three raters, rater 3 seemed to be more 'generous' in giving marks to the samples' work. Thus, the mean of the total marks for rater 3, 14.26 (57.04%), indicates that more samples obtained higher marks than lower marks. Rater 1 was seen to be the most strict rater of the three. However, she distributed marks well spread out and therefore I could describe this distribution as a near-normal one.

Immediately after the raters finished marking, they met to compare their marks and discus any difference of opinion they might have. Among the matters discussed was the difference between the three raters in giving the marks involving samples 4, 8, 9, 13, and 17. With regard to samples 4, 8, and 9, rater 1 differed from the other two raters by between 3 and 6 marks. The discussion among the raters showed that this was unavoidable. It happened because rater 1, as mentioned above, was quite strict and therefore she tended to differ from the other two between one to two marks for

every aspect of writing that had been assessed. With reference to sample 13, rater 2 described that he gave low marks to this sample. After a revision of his marking towards the sample's essay, he agreed with the other two raters that the marks given by both of his colleague were more appropriate. Even though rater 1 was labeled as the strict rater, it was not the case when she rated sample 17's essay. She gave 5 marks for the content and the organisation of ideas while the other two raters were of the opinion that this sample had slightly diverged from the topic. As a result, her total rating differs by 4 marks from the other two raters.

We also observe that the average amount that each rater's assessment deviates from the mean is small and can be calculated as follows: for rater 1, 68% of the test population would be found within 2.89 $\pm$ 1 SD, that is 15.49 to 9.71; 95% within 2.89 $\pm$ 2 SDs, that is 18.38 to 6.82; 99.7% within 2.89 $\pm$ 3 SDs, that is 21.27 to 3.93; for rater 2, 68% of the test population would be found within 4.76 $\pm$ 1 SD, that is 18.66 to 9.14; 95% within 4.76 $\pm$ 2 SDs, that is 23.42 to 4.38; 99. 7% within 4.76 $\pm$ 3 SDs, that is 28.18 to –0.38; and for rater 3, 68% of the test population would be found within 4.30 $\pm$ 1 SD, that is 18.60 to 10.00; 95% within 4.30 $\pm$ 2 SDs, that is 22.90 to 5.70; 99.7% within 4.30 $\pm$ 3 SDs, that is 27.2 to 1.40. From these calculations, we can fit for the three raters' rating 2 SDs on the (+) side of the mean which would account for 95% of this population and 1 SD on the (-) side of the mean which could account for 68% of this population. We could describe this distribution as positively skewed, i.e. towards the top than towards the bottom end of the distribution.

From the above discussion, we may stress that the differences among the raters do not affect the rating scales that have been drafted earlier. The three raters

agreed upon many things among themselves: from the content of rating scales, the division of marks to each aspect of writing being assessed, to marks given by every rater. I therefore take this as a '*consensus mark*' for each of the samples' script. I noted all the points of the discussion from the raters so that they can be used during the briefing to the essay examiners for the final test at the AIS.

4.11.2.2 Descriptive statistics of samples from Malaysia

4.11.2.2.1 The Reading Test

In this section, I will display the descriptive statistics resulting from the data sample I collected in the pilot test at both schools described earlier.

Table 4-14: Descriptive statistics for the Reading Test (Malaysia)

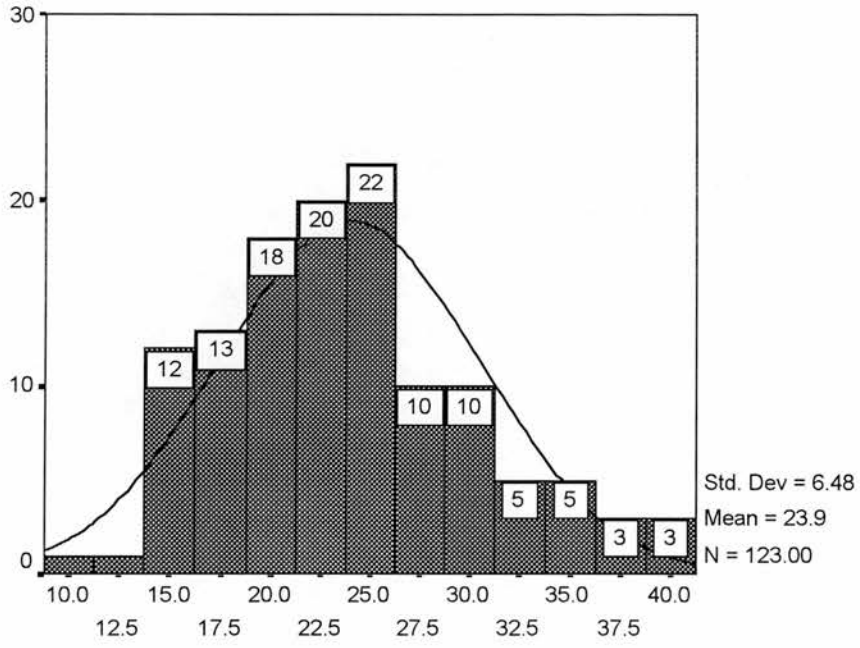
N	Valid	123
	Missing	**0**
Mean		**23.94**
Median		**23.00**
Mode		**26**
Std. Deviation		**6.48**
Variance		**41.96**
Range		**32**
Minimum		**9**
Maximum		**41**
Sum		**2945**

From Table 4-14, we note the statistical data of samples from Malaysia for the Reading Test which include central tendency and dispersion. With regard to the former, the mean, 23.94, indicates clearly that this population found the test difficult as the mean is below the 50% mark (precisely 31.92%), and there is a very high possibility that more samples are situated toward the bottom end of the distribution

than toward the top. The median, 23.00, which is slightly lower than the mean is another indication that the test was not easy for the samples. As a point below which 50% of the scores fall and above which 50% fall, the median, 23.00, indicates that the majority of the candidates obtained relatively lower marks. Another measure of the central tendency is the mode. The mode also indicates that the scores that occured most frequently were far below the 50% mark (precisely 34.66%). The last measure of central tendency is the midpoint. From the data in Table 4-14, we can calculate the midpoint for the test as 16. This is another indication to show that the majority of samples obtain relatively lower marks because the midpoint in a set of scores is the point halfway between the highest and the lowest score.

With reference to the second measure, i.e. the dispersion, we can see how widely the scores are spread out from the central tendency. The first measure is the standard deviation (SD). As shown in Table 4-14 above, the SD for this test is 6.48. From the SD and the mean of the test, I calculate the number of SDs that fit the distributions in order to describe them as normal or skewed. 68% of the test population within 1 SD would be 17.46 (negative side) and 30.42 (positive side); 95% of the test population within 2 SDs would be 10.98 and 36.90; 99.7% of the test population within 3 SDs would be 4.50 and 43.38. With the minimum score of 9 and the maximum of 41, we can fit in 2 SDs only on the negative side of the mean which would account for 95% of the test population. On the other hand, we can nearly fit 3 SDs (less 2.38) on the positive side of the mean which would account for 99.7% of the test population. We could therefore describe this distribution as positively skewed. To get a clearer picture of the distribution, I display in Figure 4-3 below the histogram of Reading Test for sample from Malaysia (N=123).

Figure 4-3: Histogram of the Reading Test (Malaysia)



The histogram in Figure 4-3 shows that the distribution is positively skewed. This confirms what has been discussed earlier, that is, more candidates scored low marks than high marks. We can see also from that the ends of the normal curve line disappear off the histogram not exactly at 0 and 41 but higher up. This confirms the calculation I made earlier concerning the 3 SDs on the negative and positive sides, that is 4.50 and 43.38.

The second statistical measure for dispersion is the variance. From Table 4-14, we can see that the Reading Test's variance for samples from Malaysia is 41.96. The variance indicates the average of the squared differences of candidates' scores from the mean. The last statistical measure is the range, i.e. the number of points between the highest and the lowest score. The range for this test as shown in Table 4-14 is 32.

4.11.2.2.2 The Grammar Test

In this section, I will display the descriptive statistics resulting from the data sample I collected in the pilot test at both schools mentioned earlier.

Table 4-15: Descriptive statistics for the Grammar Test (Malaysia)

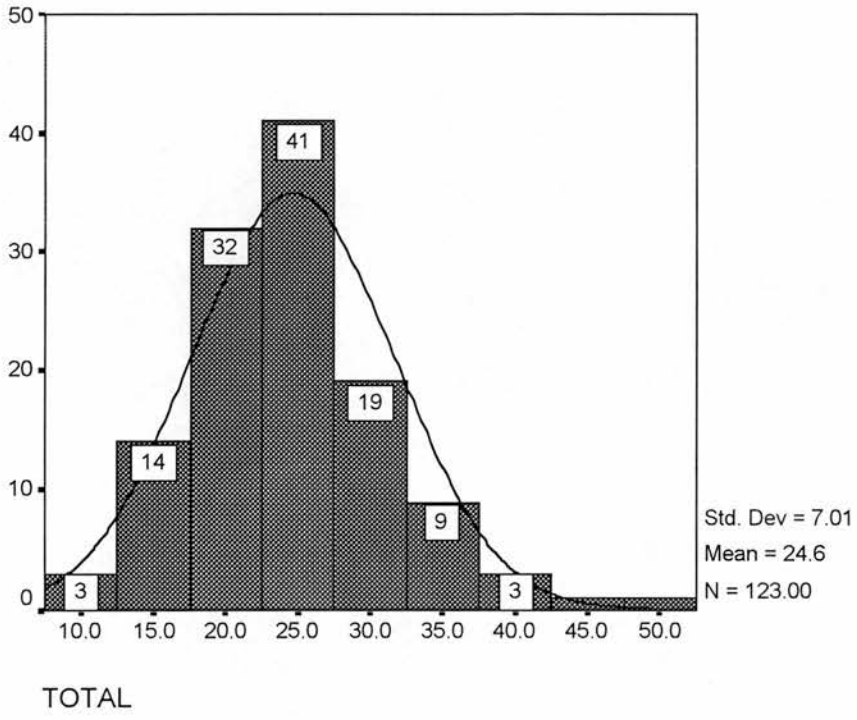
N	Valid	123
	Missing	**0**
Mean		**24.62**
Median		**24.00**
Mode		**26**
Std. Deviation		**7.01**
Variance		**49.11**
Range		**43**
Minimum		**9**
Maximum		**52**
Sum		**3028**

From Table 4-15, we observe the statistical data of samples from Malaysia for the Grammar Test which include central tendency and dispersion. As stated earlier, the total mark for the test was 60. With regard to the central tendency, the mean, 24.62 (41.03%), indicates that this population found the test quite difficult, even though they found it not as difficult as the Reading Test that has been discussed earlier. From the mean, we can say that there is a possibility that more samples are situated toward the bottom end of the distribution than the top. The median, 24.00, which is slightly lower than the mean, is another indication that the test was not easy for the samples. As a point below which 50% of the scores fall and above which 50% fall, the median, 24.00, indicates that the majority of the candidates obtained relatively lower marks. Another measure of the central tendency, is the mode. The mode, 26, also indicates that the scores that occurred most frequently were below the 50% mark (precisely 43.33%). It is odd however to find the mode here higher than the mean and the median because in the positively skewed distribution the position of the mode will normally be under both the median and the mean. The last measure of central tendency is the midpoint. From the data in Table 4-15, we can calculate the midpoint for the test as 21.5. This is another indication to show that the majority of samples obtained relatively lower marks because the midpoint in a set of scores is that point halfway between the highest and the lowest score.

With reference to the second measure, i.e. the dispersion, we observe how the scores for the Grammar Test were spread out from the central tendency. The first measure is the standard deviation (SD). As shown in Table 4-15, the SD for this test is 7.01. From the SD and the mean of the test, I calculated the number of SDs that fit the distribution. 68% of the test population within 1 SD would be 17.61 (negative

side) and 31.63 (positive side); 95% of the test population within 2 SDs would be 10.60 and 38.64; 99.7% of the test population within 3 SDs would be 3.59 and 45.65. With the minimum score of 9 and the maximum of 52, we can fit in 2 SDs only on the negative side of the mean which would account for 95% of the test population. However, we can fit 3 SDs on the positive side of the mean which would account for 99.7% of the test population. We could therefore describe this distribution as skewed towards the top of the distribution. In technical terms, this distribution is termed as *positively skewed* which means that most of the candidates scored low or average marks. To see the distribution of samples more clearly, I display in Figure 4-4 below the histogram of the Grammar Test for samples from Malaysia (N=123).

Figure 4-4: Histogram for the Grammar Test (Malaysia)



The histogram in Figure 4-4 shows that the distribution is positively skewed. This confirms what was discussed earlier, i.e. more candidates scored lower than higher marks. We can also see from Figure 4-4 that the end of the normal curve line disappears off the histogram not exactly at 0 at the negative side but higher up. At the positive side, the line disappears off just after the 50. This confirms the calculation I made earlier concerning the 3 SDs on the negative and positive sides, that is 3.59 and 45.65.

The second statistical measure for dispersion is the variance. From Table 4-15, we can see that the variance for the Grammar Test for samples from Malaysia is 49.11. The last statistical measure is the range, i.e. the number of points between the highest and the lowest score. The range for this test as shown in Table 4-15 is 43.

If we compare the data for samples from Jordan in Table 4-3 with the data in Table 4-15, we note some differences and some similarities that represent the performance of samples from both countries in the same test, i.e. the Grammar Test. For instance, the mean (37.36) for samples from Jordan was bigger than the mean for samples from Malaysia: 24.62. This is the case because more samples of the former group obtained higher marks than the samples from the latter. The standard deviation of the distribution for samples from Jordan was almost the same as that of samples from Malaysia (7.30 and 7.01). However, the total number of SDs that fit the distributions of both samples differed. This is the case because the distribution of marks for samples from Jordan was negatively skewed while the distribution of marks for samples from Malaysia was positively skewed. In non-statistical terms, more samples from Jordan obtained higher marks than samples from Malaysia. With reference to variance, there was a small difference between the variance for samples

from Jordan and Malaysia as shown in Table 4-3 and Table 4-15 respectively. If we relate the comparison of these two samples with statistical studies, we find that the above data confirms the assumption made by some statisticians that the "...greater the difference in standard deviation (or variance) between two samples, the less accurately can we establish the significance of the difference between their means" (Rowntree,1991:123). It is clear therefore from the data from both samples above that the differences between the SD and the variance are small. Therefore we can observe the big difference between the mean (37.36 or 62.3%) for samples from Jordan against the mean (24.62 or 41.0%) for samples from Malaysia.

4.11.2.3 Item analysis of the pilot test for samples from Malaysia

Three item analysis instruments will be used in analysing the test items: item facility analysis, item discrimination index analysis and distractor efficiency analysis.

(i) The Reading Test

4.11.2.3.1 Item facility (IF) analysis for samples from Malaysia

Every question in the Reading Test was computed on the samples (N=123). Table 4-16 below shows the IF for the Reading Test.

Table 4-16: Item facility for the Reading Test (Malaysia)

Item no.	Item Facility (IF)
1	.52
2	.20
3	.19
4	.26
5	.50
6	.58
7	.83
8	.46
9	.65
10	.79
11	.68
12	.88
13	.63
14	.63
15	.50
16	.39
17	.41
18	.38
19	.44
20	.89
21	.23
22	.53
23	.51
24	.68
25	.31
26	.89
27	.32
28	.33
29	.70
30	.33
31	.02
32	.11
33	.09
34	.87
35	.32
36	.21
37	.14
38	.07
39	.09
40	.12
41	.02
42	.37
43	.03
44	.11
45	.58
46	.55
47	.42
48	.03
49	.05
50	.01
51	.00
52	.01
53	.17
54	.05
55	.07
56	.31
57	.19
58	.18
59	.20
60	.42
61	.08
62	.17
63	.44
64	.25
65	.68
66	.06
67	.05
68	.02
69	.51
70	.10
71	.06
72	.04
73	.25
74	.08
75	.00

In Part One, three questions fell below the minimum index of the difficulty for an item, (i.e. .27). In Part Two, one question falls below .27. However, 34 questions have an index below .27 in Part Three. This indicates clearly that this part was extremely difficult for the population. Two questions, 51 and 75 (see Appendix A.2.2: 459, 460), have an index of .00 which means none of the population obtained the correct answers. 19 questions have an index between .01 and .10 which mean between 1 and 12 samples answered the questions correctly and 13 questions have an index between .11 and .26 which mean between 13 and 32 samples answered the questions correctly.

With reference to the difficulty of the passages in Part Three, the literature has been ambiguous. Some studies claim that a more difficult cloze test appears to correlate more highly with proficiency and criterion measures (Darnell, 1968; Oller, 1972; Carroll *et al.*, 1985). Alderson (1979) adds that the more difficult the cloze tests the better they provide a measure of proficiency. Mullen (1979) contradicts the above views by saying that an easy passage in the cloze tests provides a better prediction of ability than a difficult one. With reference to the difficulty of the passage in this part, I have to stress here that the whole text of the cloze test is not difficult because it deals with a very basic Islamic teaching in Muslim life. For example, Question 31 (see Appendix A.2.2: 459), which has an IF index of .02 only, can be classified as an easy question because the omitted word, *awfā* or *waffā*, is very familiar to students at any higher secondary level. Question 33 (see Appendix A.2.2: 459) with an IF index of .09 is another example. The omitted word, *amran* or *shay'an*, is very familiar to students in general because the whole sentence is taken from a very famous *Ḥadīth* by the Prophet Muḥammad (Peace be upon Him) and I

believe many students will have memorised that *hadīth*. The same applies to item 48 (see Appendix A.2.2: 459) which has an IF index of .03. The omitted word, *al-salām*, is very familiar to students and Muslims in general. Students not only memorise the *ḥadīth* that is related to this word, but also practice the use of *salām* (greeting each other) in their daily life. Other questions such as 32, 37, 39, 54, 60 (see Appendix A.2.2: 459-60), and the like, which have low IF indices, can be considered easy questions too. Other questions such as 38, 41, 50, 51, 52, 61, and the like are obviously difficult. Therefore it is expected that these samples will obtain low IF indices for these questions. However, many questions with varying difficulty levels and less satisfactory discriminatory values, which will be discussed later, need to be included because they are important for measuring different traits in the content domain. In other words, the decision as to whether or not to drop certain questions is often made in favour of content validity rather than test item statistics. In this regard, Heaton (1979) and Valette (1979) stress that an easy question discriminates better between poor and average students, while a difficult question is more efficient in discriminating good students from the majority of students. Therefore I intend to retain these difficult questions for the future version of the test.

There are other factors to be considered in dealing with the data in Table 4-16 above. Firstly, the samples may not be familiar with cloze tests and therefore could not perform well in this type of test. Secondly, the use of the cloze test relates to face validity. Many foreign students do not accept the cloze test as a "true" measure of their ability in language skills (Shohamy 1978; Mullen, 1979). For example, Shohamy's study on students' attitudes toward the cloze test and the oral interview revealed that only 15.38% regarded it as an accurate measure, 17% of the students

viewed the cloze as an accurate measure of speaking ability, 35.89% thought the cloze was difficult and frustrating, 10.26% claimed that it was incomprehensible, confusing, and ambiguous, and 6.41% just disliked it (quoted in Pachinburavan, 1985: 19). Taking these factors into consideration, the final decision whether or not to revise or drop items with low IF will only be made after comparing the data in Table 4-16 with the data from samples from Jordan in Table 4-4 together with the data in Table 4-17 below.

4.11.2.3.2 Item discrimination (ID) analysis for the Reading Test (Malaysia)

To obtain data for ID analysis, I employed the same procedure as was used in analysing the data for samples from Jordan. Firstly, the frequency of the Reading Test result was used to divide two groups of samples: upper and lower groups. However, the percentage of both groups was slightly higher than the percentage of both groups from Jordan: 27% (N=33). Secondly, the upper group included samples with a total mark ranging from 27 to 41. Only 1 sample (out of 4) from those who obtained 27 marks was included to make the number of samples up to 33. In the case of the lower group, 27% of it included samples with a total mark ranging from 9 to 19. Only 6 samples (out of 8) from those who obtained 19 marks were included to make the number of samples up to 33. Lastly, the ID index was calculated for each item as summarised in Table 4-17 below:

Table 4-17: Item discrimination index for the Reading Test (Malaysia)

Item	IF (upper)	IF (lower)	ID
1	.82	.30	.52
2	.42	.12	.30
3	.37	.09	.28
4	.42	.09	.33
5	.64	.52	.12
6	.82	.30	.52
7	.94	.70	.24
8	.67	.27	.40
9	.79	.67	.12
10	.91	.67	.24
11	.79	.64	.15
12	.94	.79	.15
13	.76	.49	.27
14	.61	.55	.06
15	.61	.39	.22
16	.61	.15	.46
17	.49	.33	.16
18	.39	.30	.09
19	.33	.42	-.09
20	.97	.79	.18
21	.30	.12	.18
22	.55	.39	.16
23	.61	.39	.22
24	.85	.55	.30
25	.27	.21	.06
26	.97	.85	.12
27	.33	.27	.06
28	.27	.30	-.03
29	.82	.61	.21
30	.58	.24	.34
31	.06	.00	.06
32	.27	.00	.27
33	.15	.09	.06
34	.91	.85	.06
35	.61	.15	.46
36	.36	.12	.24
37	.24	.09	.15
38	.15	.03	.12
39	.21	.00	.21
40	.27	.03	.24
41	.03	.00	.03
42	.55	.21	.34
43	.06	.00	.06
44	.27	.03	.24
45	.76	.33	.43
46	.76	.33	.43
47	.55	.09	.46
48	.06	.00	.06
49	.15	.00	.15
50	.00	.00	.00
51	.00	.00	.00
52	.00	.00	.00
53	.15	.03	.12
54	.09	.03	.06
55	.15	.03	.12
56	.49	.24	.25
57	.12	.06	.06
58	.30	.06	.24
59	.46	.06	.40
60	.73	.15	.58
61	.18	.00	.18
62	.15	.09	.06
63	.82	.12	.70
64	.42	.06	.36
65	.79	.52	.27
66	.06	.03	.03
67	.12	.03	.09
68	.06	.00	.06
69	.79	.18	.61
70	.21	.09	.12
71	.12	.00	.12
72	.12	.03	.09
73	.39	.06	.33
74	.21	.03	.18
75	.00	.00	.00

To draw conclusions about the items based on the ID indices above, I will use the guidelines suggested by Ebel (1979:267) together with the arguments by Alderson *et al.* (1996) and Brown (1996) which have been discussed earlier. With reference to Part One (Questions 1-10), 2 questions: 5, and 9 (see Appendix A.2.2: 456, 457), have ID indices below .19 which indicate that these items did not discriminate well. If we compare this finding with the finding for samples from Jordan, these questions were among 5 questions that have ID indices below .19. However, I intend to retain these questions as a good start for candidates because the IF indices for both questions, as shown in Table 4-16, indicate that these questions are at the average level of difficulty (.50, .65). Furthermore, both groups, upper and lower, also obtained relatively high IF indices. In Part Two (Questions 21-30), 13 questions have ID indices below .19 which means that these questions did not discriminate well. 6 questions(18, 19, 21,25, 27, 28) (see Appendix A.2.2: 458, 459) have low DI indices as a result of lower IF indices of both groups while the other 7 (11, 12, 14, 17, 20, 22, 26) (see Appendix A.2.2: 457-59) have similar low ID indices as a result of higher IF indices. After comparing the questions that have low ID indices in Table 4-17 with questions for samples from Jordan in Table 4-5, I conclude the following:

(A) Questions 11 and 12 (see Appendix A.2.2: 457) are found to be easy and therefore they cannot discriminate between the upper and lower groups. This was also the case with samples from Jordan in which the ID for these questions were .13 and .20 only. It is difficult to change the current questions to the new ones because the text is too short. As a way out, I will add another paragraph under the same topic from the same resource book and then make necessary changes to the questions.

(B) Questions 14 (see Appendix A.2.2: 458) is found to be easy because it cannot discriminate between both groups of samples from both nationalities. A new question will be constructed for this item.

(C) The following action has been taken with regard to questions 17, to 20 (see Appendix A.2.2: 458). With reference to questions 17 and 18, though IF indices of both groups were relatively average for samples from Malaysia, as shown in Table 4-17, it was not the case for samples from Jordan: they obtained higher IF indices ranging from .73 to 1.00. Therefore I have decided to retain both questions for the final version. With reference to questions 19 and 20, some necessary changes have had to be made. Question 19 has a negative ID index which means more samples in the lower group obtained the correct answer than in the upper group. "There is obviously something very wrong with such an item and it should be revised or discarded" (Alderson *et al.* 1996:82). I have therefore replaced Question 19 with another one even though the ID index of this question was high for samples from Jordan. Question 20 did not discriminate: the ID index was .18 for samples from Malaysia and .00 for samples from Jordan (see Table 4-5). Both groups, lower and upper, obtained high IF indices which means the question was very easy. I have therefore replaced the wording of the sentence in this question.

(D) With reference to questions 21-25 (see Appendix A.2.2: 459), I have decided to replace Question 21 with a new one. Even though Question 21 has a very high ID index for samples from Jordan (73), the IF index of this question for samples from Malaysia was very low (.23). Questions 22 and 25 are retained because the ID indices for samples from Jordan together with IF indices for samples from Malaysia were high. However, words were added to Question 22 to make it clearer for the

final test.

(E) All items in questions 26 to 28 (see Appendix A.2.2: 459) have been changed. This is due to the low ID indices of samples from both Jordan and Malaysia as shown in Table 4-5 and Table 4-17. Question 28 for example, has a negative ID index, -.03, which means more samples in the lower group were correct than in the upper group.

In Part Three (Questions 31-75), 25 questions have ID indices below .19 for samples from Malaysia. This part however will not be analysed statistically in order to drop or change certain questions as we did with items in Part One and Two. This is because the nature of questions in a cloze test are not independent from one another. Furthermore, "...items are embedded in the passage: therefore, items detected by an item analysis as not functioning properly cannot be refined in most cases" (Pachinburavan, 1985:59). The purpose of displaying the above data is to compare questions that have low ID indices with questions for samples from Jordan in Table 4-5. After comparing both sets of data, I find that only one question, 70, does not discriminate well between the upper and lower groups for both samples from both countries. Hence I have decided that no revision needs to be made on Part Three of the Reading Test. (For a new version of the Reading Test, see Appendix A.2.3: 469-481).

4.11.2.3.3 Distractor efficiency (DE) analysis

Table 4-18 below summarises the percentage of options made by the samples from Malaysia which will be used to analyse the DE for Part One of the Reading Test.

Table 4-18: Distractor efficiency (DE) statistics (N=123)

Items	Options			
	a	b	c	d
1	17	16	15	52*
2	21	20*	7	51
3	41	19*	17	21
4	67	7	26*	0
5	50*	21	19	9
6	33	2	7	58*
7	83*	3	6	6
8	46*	9	7	35
9	4	14	7	65*
10	2	2	5	79*

* correct option

To analyse the above data, I will compare them with the data in Table 4-6 for samples from Jordan. The DE statistics in Table 4-18 provide us with information about the proportion of samples who chose each of the options. Firstly, the above data shows those options selected by a majority of samples, even though they were in fact not a correct answer. Three questions, 2, 3, and 4, (see Appendix A.2.2: 456) are good examples of this. A close look into the options chosen by the samples suggests that this may happen for various reasons: the samples did not know the correct answers to the questions or the options themselves were confusing. In Question 2, for example, more samples chose option *a* rather than option *b* because they were probably misled by the word *waqt qaṣīr* in the question. In Question 3, option *a,* which the majority of samples chose, could be considered as a half-true answer. However, if the samples looked carefully at other options, they would ascertain that option *b* was the correct one as it happened to samples from Jordan: only 19.4% chose option *a* and 67.2 chose the correct option. Question 4 is another example of how the samples were fooled by the question. The use of the word *fawā'id* has lead the majority of them to choose option *a*, (*manāfi`*) regardless of

what was actually asked by the question. To conclude, I would suggest that the distractors in these questions are good because they were able to divert the samples from the correct answers. I therefore decided to retain them for the future version of the test.

Secondly, the above data reveal options other than the correct answer which do not seem to be very attractive: less than 5%. Examples of these options are option *d* in Question 4, and option *b* in Questions 6, 7, and 10 (see Appendix A.2.2: 465, 457). This was the case also with samples from Jordan (see Table 4-11). As for option *d* in Question 4, the main reason why the samples avoided it may be because they were not familiar with the meaning of the word *maghānim*. I suggest this because some samples from Jordan, 9%, did choose this option since it is more or less synonymous with the word *fawā'id*. With regard to option *b* in Questions 6, 7, and 10, I suspect that they were clearly wrong answers in the eye of the candidates. In other words, the degree to which these options attract can be said to be very low. However, since the texts for the questions were short, especially for questions 7 and 10, there were not many other alternatives that could replace those options. I therefore intend to retain them for the future version of the test.

Thirdly, the data in Table 4-18 shows us the correct options which were able to attract the majority of candidates. This can be seen in questions 7 and 10. These two questions also attracted higher percentages of candidates for samples from Jordan: 92.5% and 97% respectively. This indicates that these questions were easy in the samples' point of view. However, I do not intend to make any changes to these options because there are not many other options which could be created.

(ii) The Grammar Test

4.11.2.3.4 Item facility (IF) analysis (N=123)

Every question in the Grammar Test was computed on the samples from Malaysia (N=123). Table 4-19 shows the IF for the test.

Table 4-19: Item facility statistics for the Grammar Test (Malaysia)

Item no.	Item Facility (IF)
1	.34
2	.85
3	.66
4	.59
5	.38
6	.18
7	.72
8	.75
9	.53
10	.51
11	.49
12	.49
13	.71
14	.27
15	.29
16	.59
17	.58
18	.69
19	.13
20	.24
21	.50
22	.27
23	.70
24	.44
25	.15
26	.68
27	.18
28	.50
29	.26
30	.20
31	.08
32	.83
33	.21
34	.35
35	.40
36	.24
37	.58
38	.39
39	.43
40	.14
41	.75
42	.48
43	.15
44	.06
45	.09
46	.15
47	.11
48	.33
49	.15
50	.24
51	.86
52	.61
53	.46
54	.37
55	.42
56	.81
57	.08
58	.66
59	.70
60	.44

As in the analysis of the Reading Test, the discussion of IF analysis in Table 4-19 above will take into consideration the IF analysis of samples from Jordan discussed earlier in Table 4-7. With reference to Part A (Questions 1-50), samples from Jordan found 4 questions, 14, 40, 45, and 47 (see Appendix A.2.2: 463, 465), to be very difficult. From these samples, all four except one fell below .27 also. Even though these questions have content validity, i.e. are related to the Arabic syllabus, they seem to be difficult items. I therefore decided to revise the wording of these questions or if necessary to discard them from the final version of the test. There are other questions that fell below .27 such as 6, 19, 25, etc. However, the decision whether or not to drop these questions will only be made after the item discrimination analysis has been conducted. I cannot simply conclude, based on the IF statistics above, that an item with a low IF index, say .20 or .15, is very difficult because the academic level of the samples is lower than the target samples in the real test. Hence the discrimination index (ID) analysis which will be discussed later will focus on questions that had IF indices below .27 (15 questions altogether) and the decision will be made after that as to whether or not to discard these questions. With reference to Part B (Questions 51-60), only one question, 57 (see Appendix A.2.2: 466), had an IF index below .27. This question, 57, also had a low IF index with samples from Jordan: .10. Although the samples from Jordan obtained low IF index for this question because they provided a half-correct answer as discussed earlier, the majority of samples from Malaysia answered it wrongly (72%) and only 18% had a half-correct answer. I therefore decided to revise this question: either to change the wording of the question or reject it from the final version of the test.

4.11.2.3.5 Item discrimination analysis (N=123)

To obtain data for ID analysis, I use the same procedure to analyse the data for samples from Jordan. Firstly, the frequency of the Grammar Test result was used to form two groups of samples, upper and lower. However, the percentage of both groups was slightly higher than the percentage of both groups from Jordan, i.e. 27% (N=33). Secondly, the upper group included samples with a total mark ranging from 28 to 52. In the case of the lower group, 27% (of it) included samples with a total mark ranging from 9 to 20. Lastly, an ID index was calculated for each item as has been summarised in Table 4-20 below:

Table 4-20: Item discrimination statistics for the Grammar Test (Malaysia)

Item	IF (upper)	IF (lower)	ID
1	.64	.18	.46
2	.94	.76	.18
3	.88	.56	.42
4	.82	.49	.33
5	.18	.15	.03
6	.52	.06	.46
7	.85	.52	.33
8	.88	.30	.58
9	.58	.18	.40
10	.42	.30	.12
11	.58	.27	.31
12	.64	.21	.43
13	.88	.51	.37
14	.39	.21	.18
15	.46	.15	.31
16	.79	.39	.40
17	.85	.27	.58
18	.82	.55	.27
19	.24	.03	.21
20	.39	.15	.24
21	.72	.21	.51
22	.39	.09	.30
23	.88	.46	.42
24	.52	.27	.25
25	.30	.09	.21
26	.82	.42	.40
27	.21	.18	.03
28	.70	.27	.43
29	.49	.12	.37
30	.27	.09	.18
31	.15	.06	.09
32	.90	.76	.14
33	.46	.12	.34
34	.67	.15	.52
35	.58	.24	.34
36	.46	.21	.25
37	.94	.30	.54
38	.61	.21	.40
39	.67	.33	.34
40	.15	.09	.06
41	.79	.70	.09
42	.76	.24	.52
43	.30	.12	.18
44	.18	.03	.15
45	.18	.03	.15
46	.24	.12	.12
47	.21	.09	.12
48	.52	.15	.37
49	.27	.06	.21
50	.49	.18	.31
51	.91	.76	.15
52	.76	.46	.30
53	.53	.32	.21
54	.46	.30	.16
55	.49	.33	.16
56	.85	.72	.13
57	.03	.12	-.09
58	.73	.58	.15
59	.73	.49	.24
60	.55	.30	.25

To analyse the ID in Table 4-20, I compared the above data with IF indices discribed earlier in Table 4-19 together with ID statistics for samples from Jordan in Table 4-8. My conclusions, drawn from this comparison, are as follows:

(A) 19 questions (5, 6, 14, 19, 25, 27, 29, 30, 31, 33, 36, 40, 43, 44, 46, 47, 49, 50, and 54) (see Appendix A.2.2: 462-66), that have low IF and/or ID indices will be retained for the final version because they have high ID indices for the samples from Jordan or those from Malaysia.

(B) 5 questions (2, 32, 41, 51, and 56) (see Appendix A.2.2: 462, 464-66), which have low ID indices for both samples from Jordan and Malaysia are considered to be very easy and hence will be removed from the final version of the test.

(C) 5 questions (45, 47, 54, 55, and 57) (see Appendix A.2.2: 465-66), which have lower ID indices for either or both samples from Jordan and Malaysia were found to be very difficult. These questions will be removed from the final version of the test.

From this revision, 50 questions will be retained for the final version: 45 for Part One and 5 for Part Two (see a new revision of the test in Appendix A.2.3: 475-479).

4.11.2.3.6 Distractor efficiency (DE) analysis

In this section, I will examine the nonfunctioning and malfunctioning distractors of every option in Part A of the Grammar Test in order to revise the test for the future version. The discussion however will not take into account questions which I have decided to remove from the final version of the test. Table 4-21 summarises the percentage of options made by the samples:

Table 4-21: Distractor efficiency statistics for the Grammar Test (Malaysia)

Item	*Options*			
	a	*b*	*c*	*d*
1	1	37	25	34*
2	0	4	85*	10
3	24	66*	0	10
4	59*	21	2	15
5	17	13*	23	46
6	62	2	17	18*
7	2	12	72*	12
8	17	17	9	55*
9	32	34*	5	26
10	35*	57	6	1
11	40	49*	5	5
12	38	4	49*	5
13	2	18	72*	7
14	27*	7	28	38
15	29*	50	8	11
16	24	3	13	59*
17	19	58*	12	11
18	15	11	5	69*
19	10	33	13*	43
20	54	24*	8	13
21	14	17	49*	18
22	63	2	5	27*
23	24	70*	2	4
24	39	9	7	44*
25	15*	19	48	18
26	68*	11	4	15
27	24	29	18*	29
28	50*	22	17	6
29	34	14	24	26*
30	18	20*	6	56
31	8*	40	32	18
32	6	2	7	83*
33	8	11	21*	59
34	7	35*	40	15
35	17	20	40*	21
36	30	24*	10	31
37	31	3	3	58*
38	5	43	39*	7
39	43*	13	7	32
40	36	14*	30	14
41	5	9	75*	6
42	24	48*	9	11
43	39	15	15*	21
44	20	29	37	6*
45	9*	30	45	6
46	55	15*	12	6
47	11*	33	42	2
48	42	10	1	33*
49	15	15*	7	51
50	23	2	38	24*

*correct answer

From Table 4-21, we observe the following: firstly, the above data show those options selected by the majority of samples, even though they were in fact not correct answers (more than 45% of the samples chose wrong options). In addition to questions 14, 40, and 47 which have been analysed on samples from Jordan, other questions that led the majority of candidates to select wrong answers are 5, 6, 10, 15, 20, 22, 25, 30, 33, and 49 (see Appendix A.2.2: 462, 463-64, 466). As stated earlier, these options need to be examined to find out the element that may divert the samples from choosing the correct options. Close investigation reveals that the majority of candidates chose wrong options for different reasons. With regard to Questions 5 and 6, for example, the majority of samples chose wrong options, *d* and *a* respectively, because of carelessness. They did not read the sentence carefully, therefore they ignored the singulars which were the correct answers and opted for the plurals as their answers. A close look at Question 10 suggests that the samples were confused by the noun of *kāna* and therefore the majority chose *b* as their answer. For Questions 15 and 33, ignorance of the structure of the future tense for weak verbs was probably the reason why the majority of samples chose options *b* and *d* respectively. With reference to Question 20, the majority of the samples assumed that the word *lā* in the question was *lā al-nāhiya* (prohibition) and not *lā al-nāfiya* (denial). Therefore the majority chose option *a*. Questions 22 and 25 are other examples of carelessness on the part of the samples. As for Question 22, they knew that the appropriate *ism al-mawṣūl* for the noun of the question was *muthannā* (the dual). Therefore, more than 94% avoided options *b* and *c*. However, the majority chose option *a*, *marfū`* (nominative) being unaware that the correct option should be *d, majrūr* (genitive). The same happens to Question 25: they agreed that the noun of *aṣbaḥa* for this

question should be *marfū`* (nominative) and therefore the majority avoided options *b* and *d* which are in the accusative and genitive conditions. However, they forgot that the noun should be in the plural form and not singular as they thought. As for Questions 30 and 49, not much can be said about the samples' choices except that the options may have been selected because the samples did not know how to say the word *`asā* in *muthannā* form (for Question 33) and they did not know the root of the word *iṣṭabara* (for Question 49). For the above reasons, I intend to retain these options in the future version of the test. With reference to Question 14, however, the majority chose option *d* as did samples from Jordan. Therefore, as suggested earlier in the discussion of samples from Jordan, the word *tajāwaza* in option *d* should be replaced by another word to avoid confusion.

Secondly, Table 4-21 also reveals those options which attracted a very low percentage of candidates. Some options did not attract the samples at all (0.0%) while others had a very low percentage ranging from .8 to 10.0%. As stressed earlier in the discussion on samples from Jordan, some options cannot be easily replaced by new ones. After close investigation and scrutiny of these options, I find that options which had low percentages can be classified into two types: options that are difficult to be replaced and options that need to be replaced by new options. Options of the former type are: 6 (b); 10 (c, d); 13 (a, d); 18 (c); 20 (c); 22 (b, c); 23 (c, d); 24 (b, c); 28 (d); 30 (c); 34 (a); 36 (c); 37 (b, c); 38 (a, d); 39 (c); 42 (c); 46 (d); 49 (c); and 50 (b) (see Appendix A.2.2: 462-66). Options of the latter type are as shown in Table 4-22 below:

Table 4-22: Altered options for the Grammar Test

Items	*options*
1	a
3	c
4	c, d
9	c
16	b
17	d
25	b
26	c
33	a
47	d
48	c

It should be noted here that the modification of these options has taken into consideration the findings of samples from Jordan. With regard to Question 25 (see Appendix A.2.2: 464), even though option *b* obtained more than 10% for samples from Malaysia, I decided to replace it with a new word, *dhā*. Furthermore, this option obtained a very low percentage of samples from Jordan (3.9%).

(iii) The Dictation Test

4.11.2.4 Descriptive analysis of the Dictation Test (N=123)

As stressed earlier in Chapter Three (see 3.3.2.3.4), objective marking was used for the Dictation Test because the candidates are required to produce a response that can be marked as either 'correct' or 'incorrect'. From the text, I have selected 27 words and aspects of punctuation that will be given a mark if candidates write them correctly (see Chapter Three: 3.4.2.4.4 for the details of the words and punctuation marks which have been selected for this purpose). To start the discussion of this test, I will display the descriptive statistics resulting solely from the data samples I have

collected in the pilot test.

Table 4-23: Descriptive statistics for the Dictation Test (Malaysia)

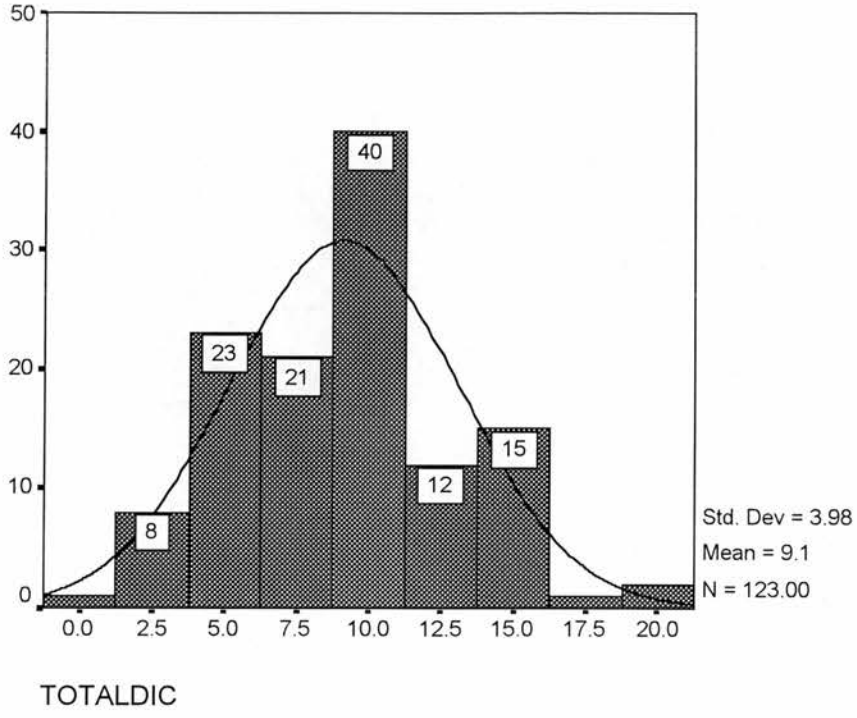
N	Valid	123
	Missing	**0**
Mean		**9.11**
Median		**9.00**
Mode		**9**
Std. Deviation		**3.98**
Variance		**15.82**
Range		**21**
Minimum		**0**
Maximum		**21**
Sum		**1121**

From Table 4-23, we note the statistical data for samples from Malaysia (N=123) for the Dictation Test which includes central tendency and dispersion. The total mark for the test was 27. With regard to central tendency, the mean, 9.11 (33.7%), the median, 9, and the mode, 9 (both 33%) indicate that the samples found the test difficult. Therefore we can predict that more samples will be situated toward the bottom end of the distribution than the top. From the maximum and minimum marks in Table 4-23, we can calculate the midpoint for the test as 10.5.

With reference to the dispersion, we can see from the data in Table 4-23 how the scores for the Dictation Test were spread out from the central tendency. With the first measure, the standard deviation (SD), we can calculate the number of SDs that fit the distribution before deciding whether or not the distribution of marks was normal or skewed. 68% of the test population within 1 SD would be 5.13 and 13.09; 95% of the population within 2 SDs would be 1.15 and 17.07; and 99.7% of the population within 3 SDs would be -.2.83 and 21.05. With the minimum score of 0 and the

maximum of 21, the calculation indicates that only 2 SDs can be fitted on the negative side while 3 SDs can be fitted on the maximum side (less .05). With these calculations, we could therefore confirm from the prediction above that many candidates scored lower marks than higher marks. This distribution, as discussed earlier with regard to the Reading and Grammar Tests, is skewed towards the top of the distribution and is termed *positively skewed*. To give a clearer picture of the distribution of marks among the samples (N=123), I display in Figure 4-5 below the histogram of the Dictation Test.

Figure 4-5: Histogram of the Dictation Test (Malaysia)



The histogram in Figure 4-5 indicates that the distribution is positively skewed. This certifies what was suggested earlier, i.e. more candidates were situated at the left side (negative) of the distribution than at the right side (positive). The end of the normal curve line, as shown in Figure 4-5 above, disappears off the histogram not exactly at 0 at the negative side but higher up. At the positive side, the line disappears off exactly at the end at the positive side. This reflects the calculation above concerning the 3 SDs on both sides, that is –2.83 and 21.05.

The second statistical measure for dispersion is the variance: the mean of the squared deviation. From Table 4-23, we note that the Dictation Test variance for this population is 15.82. This figure indicates that the marks of the population for the Dictation Test were not spread extensively and adequately. The last statistical measure is the range, i.e. the number of points between the highest and the lowest score. Since the minimum score was 0 and the maximum was 21, the calculation for the range therefore is 21.

4.11.2.5 Item analysis of the Dictation Test (Malaysia)

To analyse the Dictation Test statistically, I will employ two tools only: the item facility (IF) and the item discrimination (ID) analyses. The distractor efficiency analysis is not suitable here. The discussion starts with the first tool: the IF analysis.

4.11.2.5.1 Item facility (IF) analysis for the Dictation Test

Every item in the Dictation Test (k=27) was computed on the samples (N=123). Table 4-24 shows the IF for the Dictation Test:

Table 4-24: Item facility statistics for the Dictation Test (Malaysia)

Item no.	Item Facility (IF)
1	.41
2	.67
3	.71
4	.30
5	.68
6	.39
7	.52
8	.20
9	.14
10	.52
11	.10
12	.19
13	.13
14	.51
15	.09
16	.23
17	.008
18	.17
19	.16
20	.33
21	.72
22	.52
23	.12
24	.28
25	.59
26	.05
27	.29

From the data in Table 4-24, we observe that there is no question that had an IF index of above .90, which indicates that the question can be considered too easy. We also note that 12 questions fell below .27 (see Appendix A.2.2: 468). In other words, these questions were considered difficult for this population. These questions are related to particular aspects of the language systems as described below:

Question 8 refers to the use of *iḍāfa*;

Questions 9, 11, and 19 refer to the combination of two and three words but pronounced as one word;

Questions 12 and 15 refer to the use of *alif lam al-shamsiyya* and *alif lam al-qamariyya* respectively;

Questions 13 and 17 refer to the use of *ḥarf dhal* and *ḥarf ẓa'* respectively;

Question 16 refers to the use of punctuation comma;

Questions 18 and 26 refer to the use of long vowels; and

Question 23 refers to the use of *hamzat al-waṣl*.

From the above explanation, we could say that the samples made mistakes in various aspects of the language system when they transcribed the text. To make the

analysis easier, I divide these questions that had low IF indices into two groups: questions that have IF indices of .10 and below and questions that have IF indices above .10. Questions that belong to the former group are 11, 15, 17, and 26; and questions that belong to the second group are other than those four. The discussion, however, focuses on the questions in the first group. As for the questions in the second group, the decision whether or not to analyse them and then whether to discard or retain them for the final version, can be made only after an item discrimination analysis has been conducted.

Question 11 tested the ability of the samples to differentiate between two words which were pronounced as one word. From marking, I found that the majority of the candidates were able to separate these two words. However, they were unable to write the correct word: most of them wrote the word *man* (anyone) instead of *man`* (prohibiting). I have a feeling that this question did not achieve what it was supposed to test. Unless the ID for this question is very high, I intend to discard this question. With regard to Question 15, I noticed, during the marking too, that the majority of samples did not write the second *lam* when dictating the word *li al-khuṭūra*; most of them wrote it *li khuṭūra* instead. In addition to the above mistake, a small number of them also made a mistake when dictating *ḥarf ṭa'*, while some other samples missed the word completely. I therefore intend to retain this question for the final version of the test. With reference to Question 17, it was found, from listening to the tape, that the tester did not pronounce the item clearly. As a result, the majority of the samples were not able to dictate it properly. Therefore, I decided to omit this question from the final version of the test. However, if the ID index of this question is found to be high, the decision may be reversed. For Question 26,

observation during marking shows that the majority of the samples ignored the two letters, *waw* and *alif*, at the end of the word *ḍaminū*. This is a serious mistake showing the ignorance or carelessness of the samples towards Arabic grammar because there is no past tense in Arabic ending with *ḍamma* unless the word is attached to *waw* and *alif*. I have therefore decided to retain this question for the final version.

4.11.2.5.2 Item discrimination (ID) analysis

The same procedure was used to obtain the ID index for the Dictation Test. Firstly, the frequency of every question in the Dictation Test (k=27) was used to identify two groups of samples, upper and lower. However, the percentage of both groups was slightly lower than the percentage of both groups from the Reading and Grammar Tests using the same samples: 26% (N=32). Secondly, the upper group included samples with a total mark ranging from 11 to 21. Only 2 samples (out of 8) from those who obtained 11 marks were included to make the number of samples up to 32. 26% of samples in the lower group had a total mark ranging between 0 and 6. One sample of those who obtained 6 marks was left out to bring the number of total samples down to 32. Finally, the ID indices were calculated manually for each question as summarised in Table 4-25.

Table 4-25: Item discrimination index for the Dictation Test (Malaysia)

Item	IF (upper)	IF (lower)	ID
1	.59	.22	.37
2	.75	.41	.34
3	.94	.34	.60
4	.53	.09	.44
5	.91	.38	.53
6	.75	.09	.66
7	.78	.22	.56
8	.34	.09	.25
9	.25	.00	.25
10	.75	.22	.53
11	.22	.03	.19
12	.44	.06	.38
13	.19	.06	.13
14	.81	.28	.53
15	.28	.00	.28
16	.41	.03	.38
17	.03	.00	.03
18	.31	.06	.25
19	.38	.06	.32
20	.44	.25	.19
21	.88	.38	.50
22	.69	.28	.41
23	.31	.06	.25
24	.56	.16	.40
25	.75	.38	.37
26	.13	.00	.13
27	.53	.09	.44

From the data in Table 4-25 above, five questions, 11, 13, 17, 20, and 26 (see Appendix A.2.2: 468) have ID indices of .19 and below. All except one question (20) are among those that had low IF indices. It has already been suggested that Questions 11 and 17 should be removed from the final version of the test. This leaves three questions: 13, 20, and 26. Question 13 relates to the use of *ḥarf dhal*. Since this is the only question that tests candidates on the use of *ḥarf dhal*, I intend to retain it for the final test. As for Question 20, I decided to retain this question because its IF index is relatively high: .33. Question 26 has already been discussed above in connection with IF analysis. This question led the majority of samples to transcribe the word *daminū* wrongly. Therefore I will retain this question for the final version of the test. In addition to the above explanation regarding these questions, the academic level of the samples must also be considered. It does not necessarily follow that, because these samples obtained low marks for particular items, the candidates in the real test will obtain the same result. I believe that the candidates in the real test, based on their academic ability, can perform better not only with the overall questions but also with questions that have low IF and ID indices. (See a new version of the dictation's answer sheet in Appendix A.2.5: 505)

4.12 The time factor for the tests

As stated earlier in the pilot test administration (see 4.4), students were asked to write down on their answer sheet the time at which they finished. In this section, I will discuss the feedback from samples regarding the time limit for every test including every section of the test itself. The discussion begins with samples from Jordan,

followed by samples from Malaysia.

4.12.1 Feedback from samples from Jordan

4.12.1.1 The Reading Test

The analysis of the feedback of samples from Jordan include the following: 23 Arab samples from the Faculty of *Sharī`ah*, 21 Arab samples from the Faculty of Arts, an 33 samples from Malaysian students in Jordan. With reference to Part One (see Appendix A.2.2: 456-57), we note that only two samples finished their test after the time was up: the first was in the 11th minute and the second in the 12th minute. The rest finished before or exactly on time, i.e. within 10 minutes. Unfortunately, a large number of samples did not disclose the time they finished. I should note here that this happened due to my carelessness: I did not inform the Dean of the Language Centre at the University of Jordan to instruct samples from the Faculty of Arts to indicate the time at which they finished. As a result, none of these samples indicated in their answer papers the time at which they finished. However, I presume that the instructor had instructed the candidates for the test to stick to the time limit in the test papers. Taking this into account, I therefore include these samples under those who finished their test before or just when the time was up. In addition to these samples (N=21), a small number of samples from the other two groups did not also indicate the time they finished. With reference to Part Two (see Appendix A.2.2: 457-59), nine samples finished their Part Two after the time allocated for that part (20 minutes) was up. As for Part Three (see Appendix A.2.2: 459-60), only three samples admitted that they finished after the time was up. Oddly, a large number of samples, excluding the samples from the Faculty of Arts, did not indicate the time they finished. I noticed

that the majority of these samples obtained low marks for this part (cloze test). It is difficult to estimate the time that those samples completed their work. However, it is highly likely that these samples finished before the time was up. They may have felt bored when they were unable to answer the questions, but at the same time may have felt too guilty to indicate the time because they did not in fact finish their test. As a way out of this dilemma, they probably decided not to write the time.

4.12.1.2 The Grammar Test

The analysis of the feedback of samples from Jordan consists of the following: 31 Arab samples from the Faculty of Arts at the University of Jordan and 46 samples from Malaysian students in Jordan. As with the Reading Test, the samples from the Faculty of Arts (N=31) did not indicate the time they finished for the same reason mentioned earlier. Therefore 40.3% of the samples had no finishing time noted. With reference to the rest of the samples, we note that none of these samples finished after the time was up. In other words, the 30 minutes that was allocated for Part One in the Grammar Test was adequate (see Appendix A.2.2: 461-66). More importantly, fourteen of the samples from Malaysia were able to finish Part One in under 25 minutes. With regard to Part Two (see Appendix A.2.2: 466), all samples finished the test before or when the time was up. It is interesting to note that about 40% of the samples, excluding those samples from the Faculty of Arts, finished their work in Part Two in between 7 and 8 minutes only. To conclude, we may say that, based on the above findings, the time allocated for every part of the test is adequate.

4.12.1.3 The Essay Test

With reference to the Essay Test (see Appendix A.2.2: 467), I was informed by the Malaysian students' representative at the University of Jordan that the time allocated for the samples (N=23) was as indicated in the question paper, i.e. 30 minutes. From the essays written by the samples, the markers were satisfied that most of them included all the points required in the given topic. What can be inferred from these two statements is that the candidates were able to write the essays within the time limit.

4.12.2 Feedback of samples from Malaysia (N=123)

4.12.2.1 The Reading Test

We can summarise feedback of the samples from Malaysia as follows:

(1) We observe that only 25% (32) finished Part One when the time was up. The rest finished after 10 minutes. If we look closely, we note that the large number of samples finished Part One between the 11^{th} and the 12^{th} minute. They represent nearly 56% (68) of the total samples of 123. A very small number of samples finished the test in the 13^{th} and 14^{th} minute. They represent less than 20% of the total samples of this group.

(2) With reference to Part Two, about 56% (71) finished this part after 20 minutes and 42.3% (52) finished this part before or when the time was up. A large number of samples, 43.1%, finished this part in the 21^{st} and 22^{nd} minute.

(3) As for Part Three, 46.3% (57) finished this part before or when the time was up and about the same percentage, i.e. 46.9%, finished this part

between the 21st and 24th minute. A small number of samples did not disclose the time they finished this part.

I conclude from the above summary that relatively the time allocated for every part of the test is sufficient. Even though a large number of samples in some part, especially Part One, finished their test after the time was up, they finished it between one to two minutes only from the allocated time. Having considered the academic level of these samples, when compared to the academic level of the candidates in the real test, this small difference could be tolerated.

4.12.2.2 The Grammar Test

With regard to Part One, it is noted that about 19% (24) only finished this part after the time was up, i.e. 30 minutes. This indicates that the majority of the samples were able to finish the test before or exactly as the time allocated for this part. With regard to Part Two, the majority of the samples (68%) were able to finish this part between the 7th and 8th minute and none of them finished this part after the time was up. I recalled the suggestion by the postgraduate students earlier regarding this matter (see 4.6.1.2). They commented: "...there is no need to extend the time limit for this part [Part Two] as the matter relates to the content of the test and not to the time itself". We therefore may conclude here that, based on this finding together with the finding from samples from Jordan, the time allocated for every part of the test is adequate.

4.13 Summary of Chapter Four

This chapter has attempted to analyse the internal validity of the test items for the draft test. Three groups of samples, representing different levels of educational background, and various levels of scholastic and Arabic proficiency, participated in the analysis. They answered the questions and commented on the content and the format of the draft tests. Two major types of validity were investigated: *face* and *content*. In order to obtain these two types of validity, two major statistical tools were employed: *descriptive* statistics and *item* analyses. With the descriptive statistics analysis, the outcome of the analysis showed a very clear picture of the performance of the samples in the test: whether samples' marks centre around a particular category eg. normal, negatively or positively skewed, or whether the samples' marks are spread out from the centre. In other words, this analysis gave a picture of the distribution of marks among the samples, which helped the researcher to determine the degree of difficulty of the test. The descriptive analysis, discussed above, consisted of *central tendency* which includes the mean, mode, median, and midpoint, and *dispersion* which involves the standard deviation, variance and range. As for item analysis, three main instruments were used: item facility (IF), item discrimination (ID), and distractor efficiency (DE) analyses. Having the data from item facility analysis, we can clearly observe questions with a high or moderate or low level of difficulty. The data from IF analysis on the test questions was used as a preliminary finding before any decision could be made as to whether or not questions whose IF indices were too high or too low could be discarded. The reason was simple: questions with low IF index may discriminate well between the candidates at different levels or questions with high IF index may not discriminate well between the candidates. The process of item analysis

continued with the investigation of the data from the item discrimination index. In this analysis, we distinguished between questions with a high discrimination power and questions with a low discrimination power. Lastly, the task of item analysis was brought to its conclusion with the distractor efficiency analysis. Since the data for the IF, ID and DE analyses were obtained from different types of samples, this analysis was quite challenging. For example, questions with a high IF index for samples from Jordan may not necessarily have had the same degree of IF index as samples from Malaysia. Also, some questions with low IF and ID indices need to be retained because they are closely related to the syllabus. Some options with low percentages in the DE analysis cannot be replaced by other options because there is no suitable option available. Therefore, the decision as to whether or not to discard, to modify, or to retain questions was made after considering various factors: the comparison between the degree of the IF, ID, and DE indices of questions; the academic level of samples who took part in the pilot study (secondary school or university); the language proficiency of the samples (native or non-native speakers); and also the relationship between the test questions and the syllabus. Having considered these factors, a question may be retained for the final version if it is closely related to the syllabus even though the IF or ID indices for this question were found to be low. As a result of this analysis, modifications have been made to three sub-tests, namely the Reading, Grammar, and Dictation Tests, described in detail earlier. It is therefore anticipated that this modification will ensure the effectiveness and the usefulness of the test questions in the final test, which will be discussed in the next chapter. It is hoped that the overall findings of internal analysis in this chapter will contribute significantly, in terms of concurrent, predictive and construct validity, to the external

analysis which will be discussed in the next chapter too.

5. CHAPTER FIVE: TEST ADMINISTRATION, RELIABILITY, CORRELATION AND EXTERNAL ANALYSIS OF THE PLACEMENT TEST

5.1 Introduction

In Chapter Four, I discussed extensively the internal validity of the contents of tests, i.e. face and content validity. In this chapter, the focus of the discussion will be on the external validity which includes concurrent and predictive validity. However, the internal validity which was discussed in Chapter Four will be briefly treated in this chapter since some modifications have been made to the questions of the tests. Moreover, the items in the final version are considered the real task of the research. Therefore the effectiveness or ineffectiveness of questions which have been selected for use in the final version, based on the findings of the pilot study, will only be corroborated if an internal validation analysis has been conducted. Before investigating the external validity of the test, I will discuss the reliability of the test, in order to examine whether or not the test is consistent. Then, the correlational analysis, the last statistical tool, will be conducted. This type of statistical analysis is not less important in language testing. Brown (1996: 151) views correlational analysis as an important tool to help teachers or test makers to determine "...the degree of relationship between two sets of numbers and whether that relationship is significant (in a statistical sense), as well as meaningful (in a logical sense)". To start with, the section below describes the administration of the final version of the test which took place at the Academy of Islamic Study (AIS) at the University of Malaya, preceded by the investigation of the content validity of the final version of the test by

the Arabic language instructors.

5.2 The administration of the final test

In this section, I will describe the administration of the final version of the test. The proper administration of the test is crucially important because "...the very concept of a standardised test implies rigid control over the conditions of administration" (Clemans,1979:190). Clemans adds that although "...norms are an important part of the standardisation data, ...they will be meaningful only if derived from the administration of the test under the established conditions" (p.190). The discussion of the administration of the final version of the test, which involved running the test and marking procedures, is preceded by the investigation of content validity.

5.2.1 Examining the content validity of the test

Immediately after having altered some questions of the test, as suggested in Chapter Four, and before administering the final version to new students at the AIS, I took the test questions to a group of Arabic teachers at the university to investigate further the content validity, i.e. the *representativeness* or *sampling adequacy* of the test. In this regard, Weir (1988) in Kattan (1990:159) is of the view that "...content validity is problematic, given the difficulty in characterising language with sufficient precision to ensure the representativeness of the sample of tasks included in a test". As a way out, Weir (op. cit.) in Kattan (op. cit: 160), "...suggests a close scrutiny of the specification for a proficiency test by experts in the field and the relating of the specifications to the test as it appears in the final form". Alderson *et al.* (1996: 173) are also of the view that "...content validation involves gathering the judgement of

'experts': people whose judgement one is to trust, even if it disagrees with one's own". Alderson *et al* (op. cit.) elaborate further claiming that typically, content validation involves experts' judgement in some systematic way. "A common way is for them to analyse the content of a test and to compare it with a statement of what the content ought to be" (p.173). Alderson *et al.* suggest that this statement may be the test's specification, a formal teaching syllabus, or a domain specification. Alderson *et al.* end their suggestion by stating that better procedures for content validation would involve the creation of some data collection instrument.

Thus, I developed a questionnaire (see Appendix A.3.1: 525-530) using a four-point scale to assess the outlook of the test instructions (face validity), the texts used in the test, and the question format to four sub-tests, i.e. Reading, Grammar, Essay and Dictation. Since the format of every sub-test differs, the details of the contents of the questionnaire will be described under every section below. In addition to the assessment using the four-point scale, the respondents were also encouraged to write down their comments in the blank space (about half a page) provided at the end of the questionnaire form, especially if they chose scale 4 for the criteria assessed. The four-point scales that were used in the questionnaire were as follows:

1 = very suitable/very related/very clear/very understandable
2 = suitable/related/clear/understandable
3 = less suitable/less related/less clear/less understandable
4 = not suitable/not related/not clear/not understandable at all

The questionnaires were then distributed to two groups of Arabic language instructors in the Faculty of Language and Linguistics at the University of Malaya.

They were lecturers and teachers who are teaching Arabic at the AIS: the former group consisted of 8 lecturers and the latter consisted of 9 teachers. From these 17, 11 gave their feedback to the questionnaires, namely, 7 lecturers and 4 teachers, representing 65% of the total number of 17. The feedback was then installed in the computer using the SPSS programme. It was hoped that after obtaining the feedback from these experts, the test validity can be improved. The discussion below starts first with the instruments used for data collection and is followed by the results. The feedback on the Reading Test which includes the texts and the questions is presented and discussed first, followed by the Grammar Test, and then the Essay Test. Finally, the feedback on the Dictation Test is discussed.

5.2.1.1 Feedback on the Reading Test

The content of the questionnaire for the Reading Test can be divided into three categories (see Appendix A.3.1: 525-27):

(a) Part A refers to the first page of the question booklet. The purpose of this part was to obtain the feedback from the 'experts' as to whether or not the instructions were clear and understood by the candidates.

(b) Part B refers specifically to the texts in the Reading Test. The teachers were asked to assess every text from Part One to Part Three separately. The criteria used to assess the texts were as follows:

(i) the use of vocabulary;

(ii) the use of structures;

(iii) the suitability of the text to the students' ability;

(iv) the difficulty of the content of the text;

(v) the length of the text; and

(vi) cultural bias.

(c) Part C refers specifically to the questions in the test booklet. The teachers were asked to assess the questions from Part One to Part Three separately. The criteria used to assess the questions in these three parts were the following:

(i) the clarity of the instructions for every part of the test;

(ii) the clarity of the questions;

(iii) the relationship between the questions and the related texts;

(iv) the level of the questions, i.e. difficult or easy;

(v) the format of the questions;

(vi) the degree of familiarity, i.e. are the students usually exposed to such types of questions; and

(vii) the adequacy of time allocation

Below is the summary of the findings of the feedback from Arabic teachers at the AIS for the Reading Test:

With reference to Part A, the instruction, five teachers chose scales 1 and another five chose scale 2 from the four-point scales. Only one teacher chose scale 3. The mean for this variable was 1.64 (on a four-point scale). This indicates clearly that the majority of teachers thought that the instructions for the Reading Test were suitable, clear and understandable.

With regard to Part B, the texts used in the test, the details of the feedback by the teachers can be summarised as follows (all the means calculated below are on the four-point scale):

(i) With reference to the first criterion, the use of vocabulary, the means recorded are 1.55 for texts one, two, three and four in Part One (multiple choices (MC)); 1.73 for text one, 1.55 for texts two and three, 2.00 for texts four and five in Part Two (true-false (TF)); and 1.73 for the text in Part Three (cloze test (CT)). This apparently shows that the majority of teachers were of the view that the vocabulary in the texts were not difficult, i.e. it was at an acceptable level.

(ii) With the second criterion, the use of the structures, the means recorded are 1.55 for texts one, three and four and 1.45 for text two in Part One (MC); 1.55 for text one, 1.64 for texts two and three, 1.82 for text four and 2.18 for text five in Part Two (TF); and lastly 1.73 for the text in Part Three (CT). What could be inferred from these means is that only in one text (text five) in Part Two did the teachers think that the structure of the text was relatively not very easy even though it was not extremely difficult.

(iii) For the third criterion, the suitability of the text for the students' ability, the means recorded are 1.82 for texts one, three and four, 1.55 for text two in Part One (MC); 1.91 for text one, 1.64 for text two, 1.73 for text three, and 2.09 for texts four and five in Part Two (TF); and 1.91 for the text in Part Three (CT). What could be inferred from these means is that the majority of the teachers were of the opinion that all except texts four and five in Part Two are suitable for the candidates' ability.

(iv) For the fourth criterion, the content of the texts, the means recorded are 1.55, 1.45, and 1.64 for texts in Part One (MC); 1.55, 1.64, 1.73, 1.91,

and 2.00 for five texts in Part Two (TF); and 1.82 for the text in Part Three (CT). This again indicated that the majority of teachers were of the opinion that the content of the texts in the Reading Test are at the students' level. However, as in (iii) above, some respondents, 3 to 4, viewed that the content of texts four and five in Part Two are not very suitable for the students' level of ability.

(v) In the fifth criterion, the length of the texts, the means recorded are 1.55 for texts one and two and 1.64 for texts three and four in Part One (MC); 1.64 for text one, 1.73 for text two, 2.09 for text three, 2.18 for text four and 1.91 for text five: all were in Part Two (TF); and 2.55 for the text in Part Three (CT). The description of the means in this criterion is generally similar to what has been described above. However the mean for text five, the cloze, indicates that some respondents thought that the text, to some extent, is overly long.

(vi) For the last criterion, the cultural bias, the means recorded are 1.73 for text one, 1.64 for text two, and 1.55 for texts three and four in Part One (MC); 1.36 for texts one and text three, 1.45 for texts two and four and 1.73 for text five in Part Two (TF); and 1.55 for the text in Part Three (CT). From these means, we can conclude that the majority of teachers were convinced that because some of the texts are part of the students' Islamic culture while the others are shared by the common culture, the content of the texts did not have cultural bias. However, a small number of teachers argued, in their written comment, that the cultural bias statement itself was quite ambiguous. Therefore two of them left this

statement unanswered.

With regard to Part C, the questions employed in the test, the details of the feedback by the teachers are summarised as follows (the means calculated below are also on the four-point scale):

(i) The means recorded for the first criterion, the clarity of the instructions for every part of the test, is 1.45 for Part One (MC), 1.55 for Part Two (TF), and 1.27 for Part Three (CT).

(ii) The means recorded for the second criterion, the clarity of the questions, are 1.36 for Part A (MC), 1.64 for Part Two (TF), and 1.55 for Part Three (CT).

(iii) The means recorded for the third criterion, the relationship between the questions and the related texts, are 1.45 for Part One, 1.55 for Part Two and 1.73 for Part Three. It needs to be noted here that the respondents assessed the relationship between the questions and the text in Part Three by looking at the answers and the deletion rate (in this case, every sixth word).

(iv) For the fourth criterion, the level of the questions, the means recorded are 1.64 for Part One and 1.82 for both Parts Two and Three.

(v) For the fifth criterion, the format of the questions, the means recorded are 1.64 for both Parts One and Two and 1.82 for Part Three.

(vi) The means recorded for the sixth criterion, the familiarity of the students with the questions, are 2.00 for Part One, 1.82 for Part Two, and 1.91 for Part Three. It is unusual to find here that the mean for Part One is bigger than both Part Two and Three while, as far as I can ascertain, the students

are more familiar with such questions in the first part than with the questions in the last two. One respondent chose scale 4 for Part Three, i.e. the cloze test, which indicated that the students are not familiar at all with this type of test.

(vii) For the last criterion, the adequacy of allocation time for every part of the test, the means that were recorded are 2.55 for Part One, 2.73 for Part Two, and 2.36 for Part Three. It is obvious from these means that some respondents thought that the allocated time for every part of the test was not enough. For example, two respondents circled scale 4 suggesting that 10 minutes which had been allocated for Part One was not enough to an extreme grade and four respondents ticked scale 3 for the same part suggesting that the time, more or less, was not enough. With regard to Part Two, one respondent chose scale 4, six chose scale 3, and four chose scale 2. In Part Three, one chose scale 4, three chose scale 3, six chose scale 2 and another one chose scale 1. This indicates clearly that some teachers believed that the time allocated for the test, especially for Part Two, was not enough.

It is clear from the means above that, except for the time limit, the respondents' feedback was very positive. With reference to the time limit, the exercise confirmed that the purpose of the test was to be 'speedy'. In other words, the students should beat the time if they want to answer all the questions. Therefore, I intended to retain the time limit unchanged.

5.2.1.2 Feedback on the Grammar Test

The content of the questionnaire for the Grammar Test can be divided into two categories (see Appendix A.3.1: 528-29):

(a) Part A refers to the first page of the question booklet. The purpose of this part, is, as in the Reading Test, to obtain feedback from the 'experts' as to whether the instructions are clear and understood by the candidates;

(b) Part B refers specifically to the questions in the test. The teachers were asked to assess the questions from Part One to Part Two separately. The criteria used to assess the questions in these three parts were, with slight differences, similar to what was discussed earlier in connection with the Reading Test. Below are the criteria used to assess the questions in the Reading Test:

(i) the instructions for every part of the questions

(ii) the clarity of the questions

(iii) the relevance of the questions in relation to the syllabus

(iv) the level of the questions

(v) the format of the questions

(vi) the degree of familiarity, i.e. are the students familiar with the questions in their course of study

(vii) the adequacy of time allocation

(viii) the selection of sentences in the questions

Below is the summary of the findings of the feedback from the respondents who were asked to evaluate the Grammar Test:

With reference to Part A, the instructions, five teachers chose scale 1, three chose scale 2, two chose scale 3 and one did not disclose his or her choice. The mean for this variable is 1.55 (on a four-point scale). This indicates clearly that the majority of respondents thought that the instructions for the Grammar Test were, as with the Reading Test, suitable, clear and understandable.

With reference to Part One (questions 1 to 45) and Part Two (questions 46 to 50) of the questionnaire, I summarise the means of every criterion in Table 1 below:

Table 5-1: Means of the Grammar Test

Criteria	Part 1 (Q1-45)	Part 2(Q46-50)
the instructions for every part	1.55	2.09
the clarity of the questions	1.45	1.73
the correspondence of the questions to the syllabus	1.45	1.55
the level of the questions	1.55	1.73
the format of the questions	1.45	1.91
degree of familiarity	1.64	1.91
the length of the allocated time	1.73	2.27
the selection of sentences in the questions	1.64	1.82

From Table 5-1, we conclude that the majority of respondents chose scales 1 or 2 for the criteria stated in the questionnaire form. Some respondents gave their opinion by choosing scale 3, i.e. less suitable, less clear, less familiar etc. However, they did not represent the majority of the respondents involved in this survey. For example, in the first criterion, the instructions, three respondents chose scale 3, six

chose scale 2 and two chose scale 1. Therefore, the mean for this criteria increased to 2.09. Another criterion that has a mean of above 2.00 is the seventh, the time limit, for Part Two of the test. Four respondents chose scale 3 which represented 36.4% of the total respondents, six chose scale 2 and only one respondent chose scale 1. This means that the majority of the respondents thought that the allocated time for this part was "relatively enough". This brings us to the conclusion that no alteration needs to be made to the body of the test either on the content, the structure etc. or the format and the time limit since the feedback from the respondents was very positive.

5.2.1.3 Feedback on the Essay Test

The content of the questionnaire for the Essay Test is similar to the content of the questionnaire for the Grammar Test and can be divided into two categories: Part A refers to the instructions at the front page of the test booklet and Part B contains the criteria to be assessed by the respondents which can be summarised as follows (see Appendix A.3.1: 529):

(i) the clarity of the question

(ii) the relevance of the question to the syllabus

(iii) the level of the question

(iv) the format of the question

(v) the degree of familiarity, i.e. are the candidates familiar with the format of the question

(vi) the adequacy of time allocation

(vii) the relevance of the essay topic to the candidates' course

(viii) the interest of the candidates towards the topic of the essay

(ix) the selection of important points in the question

With reference to Part A, the instructions at the front page of the test booklet, all respondents chose scale 1 or 2 only, which indicates that the instructions are clear or very clear and understandable or very understandable. The mean recorded for this criterion is 1.55.

With regard to the criterion in Part B, the details of the outcome from the respondents' feedback are the following:

(i) for the first criterion, the clarity of the question, the mean recorded is 1.82. Only one respondent chose scale 3: the rest chose scale 1 or 2;

(ii) for the second criteria, the mean that was recorded is 1.18: all respondents chose scales 1 or 2 only;

(iii) for the third criterion, the level of the question, the mean recorded is 1.55. Even though one respondent chose scale 3, i.e. the level of the question was less applicable to the candidates, the majority (6 respondents) chose scale 1;

(iv) for the fourth criterion, the mean recorded is also 1.55, similar to the third criterion. The majority of the respondents thought that the format of the test was not unfamiliar to the students;

(v) the mean recorded for the fifth criterion, how familiar the candidates are with the format of the question, is the same as the third and the fourth: 1.55;

(vi) for the sixth criterion, the allocation of the time limit, the mean recorded is 1.91. The majority of the respondents chose scale 1 and 2: four chose

scale 1 and five chose scale 2. Surprisingly, one respondent chose scale 4 stating that the time allocated for the Essay Test was not enough at all. There was no further explanation from this respondent on the reason behind this selection in his or her questionnaire form; and

(vii) for the seventh, eighth, and ninth criteria, the means recorded are 1.55. More interestingly, the respondents unanimously chose the same scales in their questionnaire forms: six chose scale 1, four chose scale 2, and one chose scale 3.

I therefore concluded, from the respondents' feedback on the Essay Test paper, that the test has a very high face and content validity. The instructions on the front page of the test booklet were clear, the content of the test met the requirement of the syllabus, and the level of the test suited the students' ability. In addition, the allocation of the time limit for the students to write the essay was adequate, the given topic was authentic and would be interesting for the students, and, lastly, the important points that were listed in the question paper helped students to write the essay.

5.2.1.4 Feedback on the Dictation Test

The format of the questionnaire form for the Dictation Test was similar to the forms described earlier. Part A refers to the verbal instruction on the tape. In Part B, however, the assessment by the respondents was not totally based on the test paper only. Instead, the respondents were asked to listen to the tape that contained the text for the Dictation Test. Thus the respondents had to refer to two sources, the tape and the test paper, before making their assessment. Nine criteria were used for

the assessment of the Dictation Test: two of them were related to the tape and the other seven were related to the text itself (see Appendix A.3.1: 530). Below are the criteria:

(i) the clarity of the recorded voice on the tape

(ii) the relevance of the questions to the syllabus

(iii) the level of the questions

(iv) the format of the test

(v) the degree of familiarity: how familiar the candidates are with the format of the test

(vi) the length of the pauses in the recorded text

(vii) the selection of text for the dictation

(viii) the length of the text for the dictation

(ix) the suitability of the text for the candidates' ability

With reference to Part A, the instructions, the respondents' feedback indicated by the mean, 1.36, shows clearly that the respondents thought that the verbal instructions on the tape were clear, accurate, and understandable. Five respondents chose scale 1 and another five chose scale 2. However, one respondent did not disclose his or her choice.

With reference to Part B, the criteria, the findings of the feedback by the respondents are summarised as follows:

(i) for the first criterion, the clarity of the recorded voice on the tape, the mean recorded is 1.09. Five respondents chose scale 1, two chose scale 2, and one chose scale 3. Oddly, three respondents did not disclose their choice.

(ii) for the second criterion, the relationship between the test and the syllabus, the mean recorded for this criterion is 1.36. This clearly indicates that the vast majority of the respondents agreed that the content of the test was related to the syllabus.

(iii) for the third criterion, the level of the question, the mean recorded is even lower than the second criterion: 1.18. This is another indication that the majority of the respondents thought that the level of the test was suitable for the students.

(iv) for the fourth criterion, the format of the question, the mean is 1.73. The majority chose scales 1 and 2: six chose scale 1 and 4 chose scale 2. This clearly indicates that the respondents agreed with the format of the test; i.e. it will not confuse the students.

(v) for the fifth criterion, the students' familiarity with the test, the mean recorded is 1.45, lower than the third criterion. This is a clear indication that the respondents believed the students are very familiar with this kind of test. More interestingly, seven respondents chose scale 1.

(vi) for the sixth criterion, the length of pauses in the dictation text, the majority of the respondents found that the period used for the pauses was accurate and enough. Five chose scale 1 and four chose scale 2. Only two respondents chose scale 3. The mean recorded is 1.73.

(vii) for the seventh criterion, the selection of sentences, the mean drops below scale 2: 1.45. This indicates that the respondents believed the sentences used for the test fulfill this purpose.

(viii) for the eighth criterion, the length of the text in the test, the mean recorded is 1.73. Only two respondents chose scale 3; the rest chose scale 1 and 2.

(ix) for the last criteria, the suitability of the test, the mean is 1.36. It is very interesting to note that no respondent chose anything other than scales 1 and 2: seven chose scale 1 and four chose scale 2.

Summary

This part has attempted to investigate the content validity of all sub-tests from the Arabic teachers' perspective using the questionnaire forms format. The finding from their feedback regarding the instructions, the texts, and the content of the test is that there was a consensus view among these teachers towards the tests. In their view, the tests have a high content validity. Their opinion on the instructions, whether they were located on the front page of the booklet or in every section of the test itself, indicated that the instructions were clear and understandable: the means for all instructions fell at less than the 2.00 scale. Their opinions on different criteria for the texts or the test questions were very consistent: none of those criteria had a mean above the scale 3.00. More importantly, the majority of the means lay between 1.55 and 2.00 only. The teachers' judgement of the texts and questions can be summarised thus as of the right difficulty, highly related to the students' background, at a reasonable length in terms of time allocation, do have a good format, and do not have cultural bias. This finding strongly supports the analysis of content validity which was conducted earlier in Chapter Four, when I compared the content of the test with the content of the

syllabus at the AIS. We may conclude at this stage, based on these findings, that the test will achieve the purpose for which it was designed. The next section discusses the test administration to new students at the AIS.

5.2.2 The administration of the Arabic Placement Test (APT) at the AIS

The administration of the Arabic Placement Test (APT) took place on 3rd and 10th June 1998. The first day of the test was used for the Grammar, Essay and Dictation Tests and the second day was for the Reading Test. Five Arabic language teachers helped me to administer the test on the first day, and three teachers on the second day. All of the tests took place in the main lecture hall at the AIS. All candidates who took the test were new students at the AIS representing three faculties. The details of their proportion regarding the faculties is summarised in Table 5-2 below. It is interesting to note here that even though the test was originally for the use of this research, the AIS benefited from the results of the study in that its students were grouped according to the result. Therefore, during the briefing, the students were not told that the purpose of this test was for my research.

The number of students taking the test differed from the first day to the second day. For the first day of the test, only 420 took the test while in the second day, 483 took the test. However, after matching the students' names with all sub-tests, only 413 students attended the two days of the test. This means that of the 420 students who attended the first day of the test, 7 were absent in the second day and of 483 of those who attended the second day, 70 did not come for the first day of the test. I had no other choice than to reject those who did not sit for all of the test papers; including their results would spoil the data analysis such as the percentages,

means, standard deviations, reliability, correlation, etc. However, all papers were marked by the examiners for the purpose of allocating the students to appropriate groups for the Arabic course at the AIS.

Table 5-2: Summary of the students' frequency and percentage according to their faculty

Faculty	Frequency	Percent
Uṣūluddīn	120	29.1
Tarbiya	99	24.0
Sharī`ah	194	47.0

Two groups of examiners helped me in marking the answer papers: the first group consisted of eight Arabic teachers at the Faculty of Languages and Linguistics at the University of Malaya, who were my colleagues when I was there ten years ago, and the second group was thirty final year students at the university, who were my students at the Faculty of Education before I left the country for my study. The former helped me in marking the Essay Test (subjective test) and the latter helped me in the rest of the papers, i.e. the Grammar, Dictation and Reading tests (objective test).

5.2.2.1 The administration of marking the subjective test

With the first group of markers, the Arabic language teachers, I coordinated a *standardisation* meeting. The purpose of this meeting was to organise standard marking of the papers and familiarise the examiners with the rating scale. Having familiarised myself with the rating scale, as was discussed in Chapter Four, I personally acted as the chief examiner. With regard to the length of the meeting, Alderson *et al.* (1995) suggest that the meeting should take a complete day at least, to ensure enough discussion for all examiners. However, in a small testing

programme like the one I had, I decided that one whole day was not needed. Instead, I decided to have a very short meeting with these examiners and indeed the meeting ran for about two hours and half only. Two days before attending the meeting, I gave the examiners two scripts of essays by the samples from Jordan and the rating scale (see Chapter Four, 4.11.2.1 (iii) for details). The first script represented an *adequate* performance and the second represented an *inadequate* performance as examined earlier by the *standardisation committee* (see Chapter Four). The purpose was merely to have them try out the rating scale when marking the two scripts. They were reminded to be prepared to explain their marks to their colleagues at the meeting.

On the day of the standardisation meeting, the first stage was devoted to discussing the two scripts. It was observed, from the discussion, that two examiners gave marks to the inadequate script with a difference between 3 and 4 marks from the other examiners. (The other six raters gave the total marks to that script ranging from 9 to 10 only). The discussion revealed that this disagreement arose from an unclear conception of the rating scale by the two examiners concerned together with a 'generous' attitude of not wishing to give too many lower marks. This disagreement had been tackled by emphasising that the purpose of the meeting was to help the examiners to match their marks with the standardising committee. Both examiners were implicitly reminded that a major disagreement will affect the analysis of data and may lead to errors in the interpretation of the data. They were then advised to reduce the marks for the inadequate script so that their marks matched, more or less, the judgement of the others and the committee. With reference to the adequate script, the examiners tended to agree on the marks they had given. The

differences among them were small, ranging between one and two marks only.

The second stage of the meeting was having further practice in marking. Every examiner was given two essay scripts: one adequate, and one inadequate, by the candidates whose scripts I had photocopied before the meeting. The purpose was again to see whether or not the marking matched the judgement of the committee. The results were very consistent: five examiners reached the same mark on the adequate and inadequate scripts and the other three differed by one to one and a half marks only. This may have happened because many problems had been resolved in the standardisation meeting discussed earlier. Another reason could be that the examiners had already become familiar with the writing style of the two candidates. Since there was no suggestion of making any changes to the rating scale, we agreed to use the rating scale that had been used for practice during that meeting. At the end of the meeting, I distributed the scripts to the examiners: everyone had to mark about sixty scripts. They were given three days only to mark the scripts because the tutorials started in the third week. This was because by that time, new students needed to be informed about which Arabic group they had been assigned to. Apart from this marking session, three examiners were asked to mark 10 scripts from another examiner for the purpose of the Reliability analysis which will be described later. This exercise, however, was undertaken after the scripts were marked.

After receiving the marks from the examiners, I entered them into the computer using the SPSS package for data analysis. At the same time, I added up those marks with marks from the other three sub-tests for the purpose of grouping students according to their results. For those who attended the test for one day only, the groups assigned for them relied only on that particular result regardless of the

test that they had missed. Students who attended only one test were assigned to a group on the basis of the results of that test. This is to say that there was no second test administration for those who were absent from any test.

5.2.2.2 The administration of marking the objective tests

Three days after the second test was conducted, I had a one day 'marking scripts meeting' with thirty final year students (hereafter called the markers), in one of the lecture halls at the Faculty of Education. This place was chosen primarily because all scripts were kept in my room at the Faculty. In addition, all lecture halls in this faculty are equipped with such facilities as overhead projectors (OHP), computer terminals and extensions, etc, which I needed for this kind of meeting. The first stage of the meeting was devoted to explaining the marking schemes for the three papers: Grammar, Reading and Dictation. I used transparencies and an OHP to explain to the markers the marking schemes of these papers. The Reading Test marking scheme was explained first, followed by the Grammar Test, and then the Dictation Test. With reference to the Reading and Grammar Tests, the markers were asked to count the correct answers and to mark zero for unanswered questions. They were not allowed to put any mark or sign for wrong answers. With particular reference to Part Three in the Reading Test, the cloze, there was more than one correct answer to some questions. Therefore, I displayed the correct answers on the screen, so that the markers would be guided on how to mark this part. I obtained some additional correct answers from the samples in the pilot study, especially the samples from Jordan. (see Chapter Three, 3.4.2.1.4 and 3.4.2.2.4 for the details of the procedures of marking).

With regard to the Dictation Test, the procedure of marking the scripts was very complex, especially at the beginning of the marking session. The problem usually occurred when the candidates did not dictate certain sentences, which caused the markers to lose the sequences of the questions. This was not to include other problems like bad hand writing or dictating, which were unconnected to the test task. On many occasions, I had to interfere in the marking process, especially when the markers lost the number of the questions. The task of the markers became more difficult when I instructed them that even any small mistake should be considered an error. This was totally different from the Reading Test, especially for Part Three, the cloze, where some small mistakes like ignoring *alif lam* in definite words was tolerated.

The last stage of the meeting was devoted to counting the correct answers given by the candidates for every part of the answer scripts, and then adding up the grand total of every script. I then took all the scripts to the computer room in the Faculty for data installation. In addition to the total correct answers for every part of the test papers as well as the grand total marks, I had to enter into the computer all options made by the candidates for all of the test papers. With help from the computer assistant at the Faculty, I finished this job at the end of September 1998.

5.2.3 The descriptive analysis of the final version of the tests

In this section, I will describe briefly the data of the descriptive analysis of the four sub-tests. The purpose of this analysis is to see the central tendency and dispersions of the test. Knowing about these data helps us to compare them with the data in Chapter Four and then to conduct the analysis in the context of the entire

score distribution. One null hypothesis (H_0) and one alternative hypothesis (H_1)have been set up:

The null hypothesis is:

> There is no difference between the results of samples in the pilot study and the samples from AIS in terms of central tendency and dispersion.

The alternative hypothesis is:

> There is a difference between the result of samples of the pilot study and the samples from the AIS in terms of central tendency and dispersion.

The purpose of setting up these hypotheses is to prove, in the following descriptive analysis, that there is a difference between the two groups of samples, in terms of result, after the modifications have been made to some items. It is also to prove that any changes and differences that occurred did not happen by chance. Having proven these matters means we reject the null hypothesis. However, if these two matters cannot be proven, we therefore have to accept the null hypothesis and consequently reject the alternative hypothesis. The focus of the discussion is on the mean for central tendency and the standard deviation for dispersion.

5.2.3.1 The Reading Test (N=413)

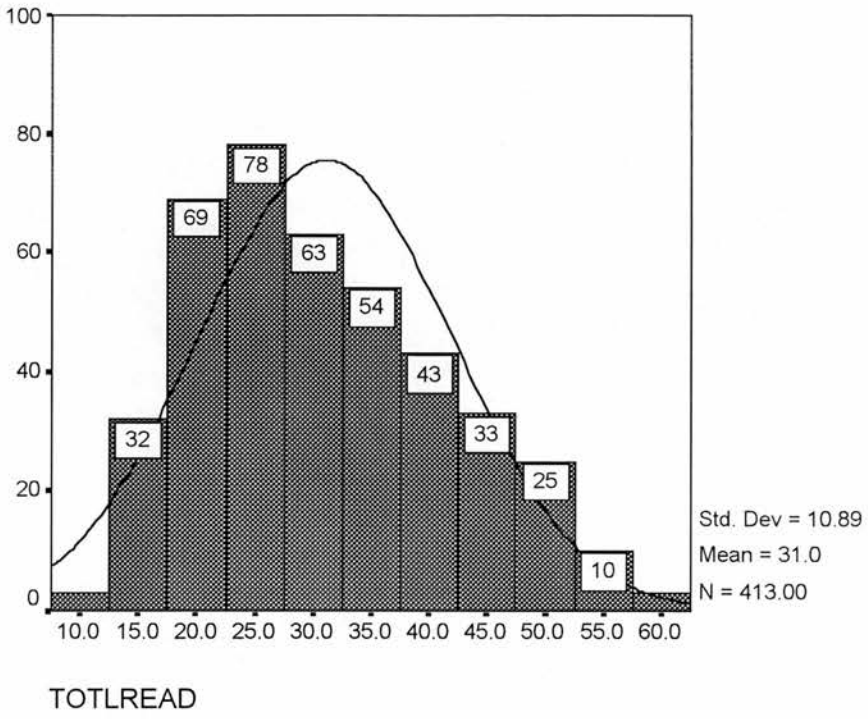
Table 5-3 below summarises the statistical data of samples from AIS for the Reading Test which include central tendency and dispersion.

Table 5-3: Descriptive statistics for the Reading Test (N=413)

N	Valid	413
	Missing	0
Mean		31.0194
Median		29.0000
Mode		27.00
Std. Deviation		10.8864
Variance		118.5142
Range		50.00
Minimum		9.00
Maximum		59.00
Sum		12811.00

The mean, 31.01, indicates that the candidates found the test moderate. There is a high possibility that the samples will be divided evenly into the normal distribution, i.e. 3SDs on both sides, top and bottom. In the pilot study, however, the total mean for the samples from Jordan (N=67) was 48.40, i.e. more samples were situated towards the top, while the total mean for the samples from Malaysia (N=123) was 23.94, i.e. more samples were situated towards the bottom. From Table 3, we note that the standard deviation (SD) for the Reading Test is 10.89. The calculation of the SDs which will fit the distribution, from the mean and the SD, indicates that we can fit nearly 3 SDs on the positive (+)(41.89, 52.78, 63.67) and 2 SDs on the negative sides (-)(20.11, 9.22, -1.67) which would account for nearly 99.7% and 95% respectively of this population. Therefore, we could describe this distribution as positively skewed. In the pilot study, the distribution of the SD for the samples from Jordan was negatively skewed, i.e. 1 SD on the (+) side (68%) and 3 SDs on the (-) side (99.7%), while the distribution for the samples from Malaysia was positively skewed, i.e. nearly 3 SDs on the (+) side (99.7%) and 2 SDs on the (-) side (95%). To get a clearer picture of the distribution of samples from AIS for the Reading Test, I display the histogram in Figure 5-1 below:

Figure 5-1: Histogram of the Reading Test (N=413)



The histogram in Figure 5-1 shows that the distribution is positively skewed. We can see that the ends of the normal curve line disappear off the histogram not exactly at 0 and 60 but higher up. This confirms the number that fit the distribution I calculated above concerning the 3 SDs, i.e. -1.67 to 63.67.

5.2.3.2 The Grammar Test

The descriptive statistics for the Grammar Test are summarised in Table 5-4

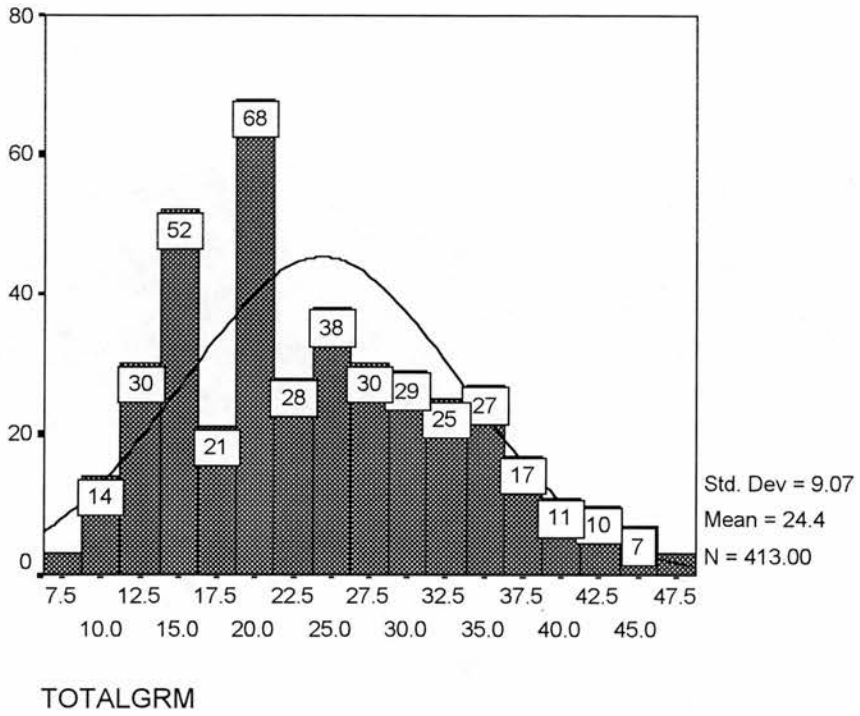
Table 5-4: Descriptive statistics for the Grammar Test (N=413)

N	Valid	413
	Missing	0
Mean		24.3680
Median		22.0000
Mode		20.00
Std. Deviation		9.0739
Variance		82.3351
Range		41.00
Minimum		7.00
Maximum		48.00
Sum		10064.00

From Table 5-4, we find that the total mean for the Grammar Test is 24.37. There is a high possibility that the distribution of the marks will be a normal one because the mean, 24.37, is equal to 48.74%. It should be noted here that in order to compare the mean for the samples from the AIS with the samples in the pilot study, a percentage will be used because the number of questions in the final version has been reduced from 60 to 50. The SD for this test, as shown in Table 5-4, is 9.07. The calculation of the SDs which will fit the distribution, from the mean and the SD, indicates that we can fit nearly 3 SDs on the positive (+) side (33.4, 42.4, 51.4) and 2 SDs on the negative (-) side (15.4, 6.4, -2.6) which would account for nearly 99.7% and 95% of this population respectively. Therefore, we could describe this

distribution as a positively skewed. In the pilot study, the total means of the samples from Jordan and Malaysia for this test were 37.36 (74%) and 24.62 (47%) respectively. This, as stated earlier in Chapter Four (see 4.11.1.1.2 and 4.11.2.2.2), indicated that more samples from Jordan were situated at the top (+ side) than the bottom end (- side). On the other hand, more samples from Malaysia were situated at the bottom end (- side) than the top (+ side). The SD for the samples from Jordan was 7.30 and the SD for the samples from Malaysia was 7.01. The distribution of marks for the samples from Jordan was negatively skewed, i.e. 2 SDs on the (+) side (95%) and 3 SDs on the (-) side (99.7%), while the distribution of the marks for the samples from Malaysia was positively skewed, i.e. 3 SDs on the (+) side (99.7%) and 2 SDs on the (-) side (95%). Figure 2 below displays the distribution of the population for the Grammar Test at the AIS.

Figure 5-2: Histogram of the Grammar Test (N=413)



From Figure 5-2, we observe that the distribution of samples in the test is positively skewed. We can see that the ends of the normal curve line disappear off the histogram not exactly at 0 and 60 but higher up. This confirms the number that fit the distribution I calculated above concerning the 3 SDs, i.e. -2.6 to 51.4.

5.2.3.3 The Essay Test

The descriptive statistics of the test are summarised in Table 5-5.

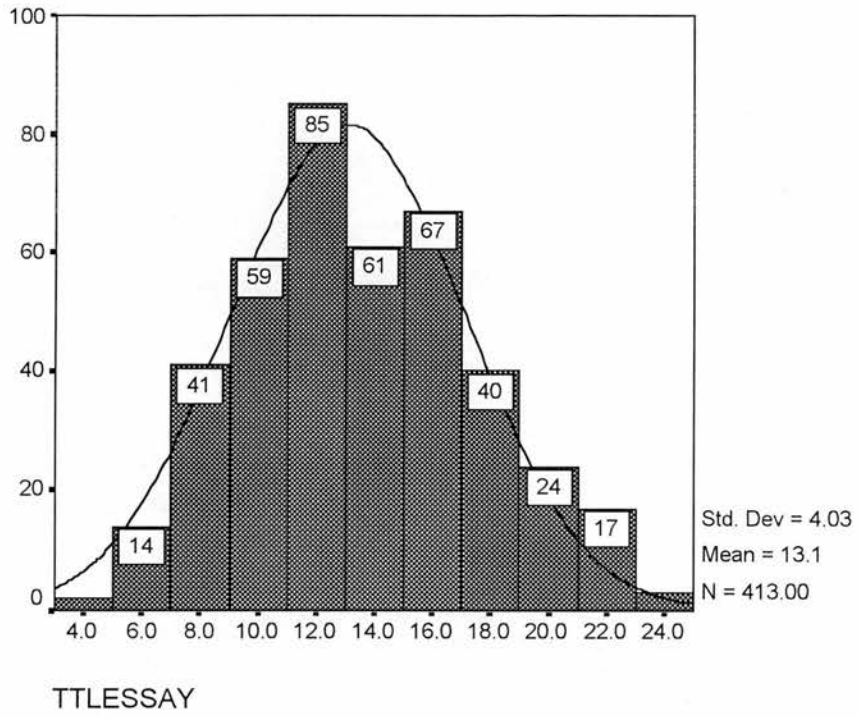
Table 5-5: Descriptive statistics for the Essay Test (N=413)

N	Valid	413
	Missing	0
Mean		13.06
Median		13.00
Mode		12
Std. Deviation		4.03
Variance		16.26
Range		20
Minimum		4
Maximum		24
Sum		5393

The mean for the test, as shown in Table 5-5, is 13.1 (54%) from the total marks of 5393. From this mean, it is highly suspected that the distribution of the population for both sides will be balanced. In other words, no samples were situated more on the top than the bottom or vice versa. The SD for the test is 4.03. The numbers of SDs that fit the distribution are 17.04, 21.07, and 25.1 for the (+) side and 9.07, 5.06, and 1.03 for the (-) side. With the minimum and maximum marks of 4 and 24, we can fit nearly 3 SDs on both sides of the mean. This distribution is clearly a normal distribution. No comparison could be made with the samples from the pilot study because there was no total score of the pilot study for the Essay Test available. However, the above finding is very meaningful because the distribution of marks was

obtained by using the rating scale which has been accepted as the *standardised* rating scale and was not obtained, in any way, from the examiners' own rating. Figure 5-3 below displays the distribution of the population for the Essay Test:

Figure 5-3: Histogram of the Essay Test (N=413)



5.2.3.4 The Dictation Test

The descriptive statistics of the test are summarised in Table 5-6 below.

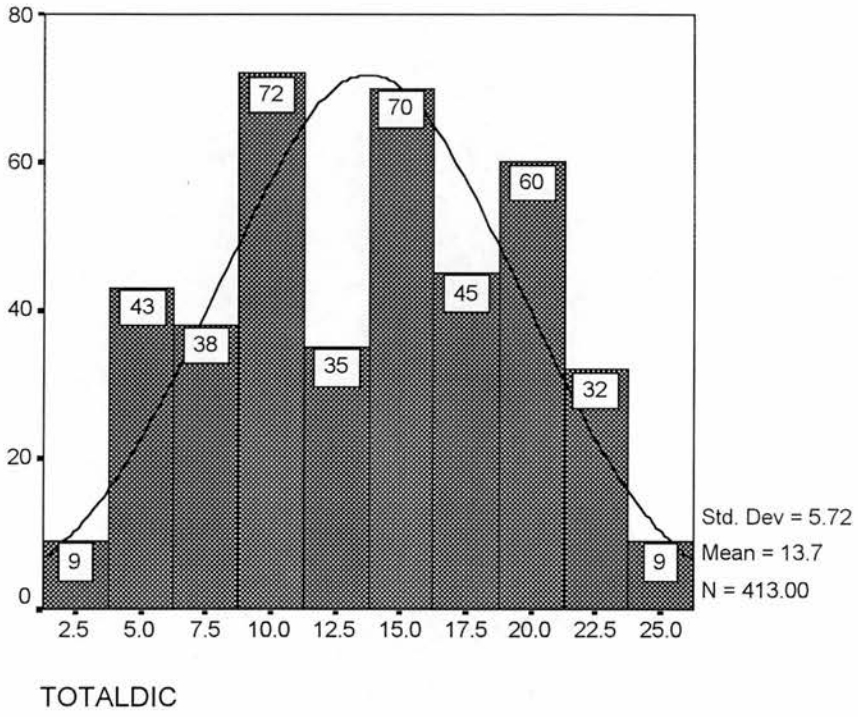
Table 5-6: Descriptive statistics for the Dictation Test (N=413)

N	Valid	413
	Missing	0
Mean		13.7215
Median		14.0000
Mode		20.00
Std. Deviation		5.7177
Variance		32.6917
Range		23.00
Minimum		2.00
Maximum		25.00
Sum		5667.00

The total mark for both tests, i.e. the Essay and the Dictation, is the same: 25. The mean for the Dictation Test is 13.72 (55%) from the total marks of 5667, slightly higher than the Essay Test. From this mean, it is highly suspected that the distribution of the population for both sides, positive and negative, will be balanced. The SD for the test is 5.7, slightly higher than the SD for the Essay Test too. The numbers of SDs that fit the distribution are 19.42, 25.12, and 30.82 for the (+) side and 8.02, 2.32, and -3.38 for the (-) side. With the minimum mark of 2 and the maximum of 25, we can fit only 2 SDs on both sides of the mean. To compare this finding with the finding of the pilot study, the same procedure, as in the Grammar Test above, will be employed because two items were omitted from the final version of the test. In the pilot study, the mean for the test was 9.11 (43%) which means that more samples were situated towards the (-) side than the (+) side of the distribution. With reference to SD, the SD for the test in the pilot study was 3.98.

The distribution of marks of the Dictation Test for the pilot test was positively skewed because only 2 SDs could be fitted on the (-) side while 3 SDs could be fitted the (+) side. This distribution is better than the distribution of marks for the pilot study, as shown by the mean of the total score above. Figure 5-4 below displays the distribution of the samples at the AIS for the Dictation Test:

Figure 5-4: Histogram of the Dictation Test (N=413)



It can thus be concluded that the modifications incorporated in the final version of the test have made some changes, in general, to the central tendency and the dispersion of all of the sub-tests. The distribution of samples on both positive and negative sides becomes more balanced where this kind of distribution was not seen with the samples in the pilot study. We are therefore obliged to reject the null hypothesis set above and consequently accept the alternative hypothesis, i.e. there is a difference between the results of samples of the pilot study and the samples from the AIS in terms of central tendency and dispersion.

5.2.4 Statistical analysis of the final test

In this section, I will describe, in brief, the statistical analysis of the final version of the tests. The analysis, however, does not include the Essay Test because the nature of this test is not suitable for statistical analysis. Three factors will be used for this analysis: item facility (IF), item discrimination (ID) and distractor efficiency (DE). I have to emphasize here that the analysis focuses more on the items that had been modified as a result of the statistical analysis in the pilot study. However, the details of the frequency of every item of the tests for every factor, i.e. IF and ID, are attached in the Appendix B.1: 536-539 for self-investigation. The analysis of the Reading Test is first presented, followed by the Grammar and then the Dictation Test. With regard to the analysis factors, the IF is presented first, followed by the ID and then the DE where applicable.

5.2.4.1 The Reading Test

(i) Item facility analysis

Several questions in Part Two of the Reading Test have been modified (see Chapter Four (4.11.2.3.1 for details). These questions are: 11, 12, 14, 19-22, and 25-28 (see Appendix A.2.3: 472-74). Some questions which were not modified such as the cloze test will be observed too. Below are the IF for questions that have been modified together with the IF indices from both samples in the pilot study for the purpose of the comparative study (see Appendix B.1: 537 for the details of the IF indices for the Reading Test):

Q no.	IF(AIS)	IF(Jordan)	IF(Malaysia)
11	.81	.73	.68
12	.90	.70	.88
14	.61	.63	.63
19	.41	.72	.44
20	.94	.93	.89
21	.24	.64	.23
22	.57	.69	.23
26	.90	.93	.89
27	.42	.54	.32
28	.39	.46	.33

From the data, I draw the following conclusion:

(a) Questions 11 and 12 have IF indices that are similar to IF indices of both samples in the pilot study. This indicates that an attempt to make the questions more difficult by adding another paragraph was not successful.

(b) Questions 14, 19, 20, 21, and 26 have IF indices that are similar to both IF indices of the samples in the pilot study. This indicates that replacing the

questions did not make a big difference to the items of the test: they remained difficult despite the modification.

(c) Questions 22, 27, and 28 have IF indices that differ from IF indices of both samples in the pilot study. This indicates that the modification of the questions had made some improvement to the IF indices of the questions of the test.

With reference to Part Three (see Appendix A.2.3: 474), the cloze test, 25 questions were found to have IF indices below .27, less 7 items from the samples from Malaysia (N=123). This part can still be considered difficult for the population.

(ii) Item discrimination (ID) analysis

The ID analysis focuses on the same questions that have been described above. The following table is the summary of ID indices for the samples from the AIS together with the indices of both samples from the pilot study for the purpose of the comparison study.

Q no.	**ID(AIS)**	**ID(Jordan)**	**ID(Malaysia)**
11	.27	.13	.15
12	.13	.20	.15
14	.45	.14	.06
19	.17	.40	-.09
20	.13	.00	.18
21	.22	.73	.18
22	.10	.54	.16
26	.25	.13	.12
27	.25	.00	.06
28	.19	-.06	-.03

From the data above, we draw the following conclusion:

(a) Four questions, 12, 19, 20, and 22 (see Appendix A.2.3: 472-73), had ID indices

below .19, the minimum limit for every question to be considered discriminating. This indicates that replacing questions with new ones did not have any strong effect. Surprisingly, the ID indices for Questions 12, 20, and 22 were lower than the ID indices in the pilot study.

(b) Six questions had ID indices above .19 which were considered discriminating. More importantly, some of these questions, such as Questions 26, 27, and 28 (see Appendix A.2.3: 474) have increased their ID indices drastically as a result of being replaced by new questions.

With reference to Part Three, the cloze test, 9 questions(20%) had ID indices below .19, i.e. did not have the minimum level of the discriminating power, less 16 questions from the samples from Malaysia (N=123). These questions are Questions 41 (.11), 43 (.04), 49 (.09), 50 (.18), 51 (.15), 52 (.13), 66 (.03), 67 (.02) and 68 (.09) (see Appendix A.2.3: 474). We may conclude here that some questions in Part Three remained difficult for this population.

5.2.4.2 The Grammar Test

(i) Item facility analysis

As suggested in Chapter Four (see 4.11.2.3.4), 10 questions with low IF and ID indices have been removed from the final version of the test. The analysis here investigates other questions that have low IF indices. From the summary of the items statistics, some questions were observed to have IF indices below .27. The Table below summarises the IF indices of these questions together with the IF indices of the samples from the pilot survey for the comparison study (see Appendix B.1: 538 for the details of the IF indices for the Grammar Test):

Part A

IF(AIS)	IF(Jordan)	IF(Malaysia)
.26 (4)	.38 (5)	.13 (5)
.16 (13)	.26 (14)	.27 (14)
.20 (19)	.42 (20)	.24 (20)
.19 (26)	.31 (27)	.18 (27)
.26 (30)	.47 (31)	.08 (31)
.26 (31)	.44 (33)	.21 (33)
.26 (34)	.62 (36)	.24 (36)
.16 (41)	.51 (44)	.06 (44)
.23 (42)	.34 (46)	.15 (46)

Note: the numbers in brackets represent the number of the question

The low IF indices of the questions above indicate that the candidates from the AIS found these questions difficult. No question in Part Two had an IF index below .27 (see Appendix A.2.3: 475-79).

(ii) Item discrimination (ID) analysis

The investigation of the ID indices for the candidates from the AIS shows that all except one question had an ID index of .19 or above for both Part One and Part Two. This shows that questions in the Grammar Test discriminated well.

(iii) Distractor efficiency (DE) analysis

As mentioned in Chapter Four (see 4.11.2.3.5), 11 options were found to have very low percentages due to various reasons which were discussed in detail. These options were replaced by new ones for the final version of the test. In this section, I display the percentage of the options which were chosen by the candidates at the AIS to see whether or not this replacement improved the effectiveness of these options to attract the candidates. The data from the pilot study is also included for the comparative study.

Options	%(AIS)	%(Jordan)	%(Malaysia)
(a)	3.9(1)	0.0(1)	0.8(1)
(c)	1.9(2)	0.0(3)	0.0(3)
(c)	2.7(3)	1.3(4)	1.6(4)
(d)	4.1(3)	0.0(4)	15.4(4)
(c)	1.7(8)	2.6(9)	4.9(9)
(b)	2.7(15)	1.3(16)	3.3(16)
(d)	5.1(16)	1.3(17)	10.6(17)
(b)	13.1(24)	3.9(25)	17.9(25)
(c)	1.9(25)	0.0(26)	15.4(26)
(a)	3.9(31)	0.0(33)	8.1(33)
(c)	3.4(43)	0.0(48)	0.8(48)

We observe from the data above that although the new options had increased the percentages of the particular options, the outcome was not encouraging. The percentages of these options were still small compared with the percentages of the pilot study. With reference to some questions such as 3, 16, and 25 (see Appendix A.2.3: 475-77), the new options attracted the candidates less than the options that were used with the samples in the pilot study, especially with the samples from Malaysia. I can conclude here that the replacement of new options, in general, did not attract the candidates.

5.2.4.3 The Dictation Test

(i) Item facility analysis

Two questions were found to have IF indices below .27: Question 12 (.24) and Question 24 (.12) (see Appendix B.1: 538 for the details of the IF indices). In the pilot study, these two questions were among five questions that had an IF index of below .27.

(ii) Item discrimination analysis

The investigation of the ID indices for the candidates from the AIS shows that no question had an ID index below .19, even with Questions 12 and 24 which had been described above as having low IF indices. Both questions had the ID indices of .55 and .33 respectively.

It may be concluded from the above discussion that the modifications incorporated in the questions of the tests have, generally, improved the IF indices and the ID indices of the items. The improvement happened either as a result of changing the wording of the questions or by replacing a question with a new question or by discarding particular questions from the final version of the test. If we compare this finding with the findings of the pilot study, we note that the final test has fewer questions with low IF and ID indices. However, some questions, as discussed above, remained difficult despite the modification. We may need to refine these questions in future in order to increase the IF and ID indices of these questions. There is an exception however: the distractor efficiency analysis of the Grammar Test shows that not many improvements can be obtained despite the replacement being made. As was stressed earlier in Chapter Four, there was no absolute guarantee as to whether or not the new options could attract the candidates to choose them: in many cases it was very difficult to find a suitable replacement for the options.

5.3 Reliability analysis of the final version of the test

The literature on the reliability coefficient (r_{xx}) of the test has been discussed

earlier (see Chapter Two, 2.3). In this section, I carry out a reliability analysis which will show whether or not the questions in the test are consistent as well as to what extent they contribute to the test's internal reliability.

There are three common ways of estimating reliability: test-retest, equivalent forms, and internal or inter-item consistency (Crocker and Algina, 1986; Anastasi, 1988; Brown, 1988, 1996; Alderson *et al.* 1996). The internal consistency reliabilities use various methods of estimating the reliability of a test. They are the split-half method, Kuder-Richardson formula 20 and 21 (K-R20) (K-R21), and Cronbach Alpha. I chose the third way of estimating reliability, the internal consistency reliabilities using the Cronbach Alpha method because "...they [internal consistency reliabilities] have the distinct advantage of being estimable from a single form of a test administered only once - in contrast to test-retest and equivalent forms of reliabilities, which require either two administrations or two forms" (Brown,1988:99). The Cronbach Alpha method was used because all except the Essay Test items were equally weighted and were marked as either right or wrong. With regard to the Essay Test, the consistency of the markers is obtained by calculating the Rank-Order Correlation of the markers. Brown (op. cit.), however, argues that the choice of methods of measuring reliability is not the main issue of the reliability analysis because regardless of the type of reliability involved, the interpretation of the coefficients will be about the same. Brown further stresses that we should be concerned "...with how consistent the test [is] in terms of the percentage of the reliable variation in scores, as opposed to error. If r_{xx} = .30, then 30 percent of the variation is reliable and the remaining 70 percent is error. [This indicates] that the test is not reliable and that another one should have been used" (p.100).

With regard to the reliability coefficient (r_{xx}) of the test's consistency, Lado (1961) and Alderson *et al.* (op. cit.) are of the view that this depends on many factors such as the type and length of the test. Lado (in Hughes 1992) for example suggests that "...good vocabulary, structure and reading tests are usually in the .90 to .99 range, while auditory comprehension tests are more often in the .80 to .89 range. Oral production tests may be in the .70 to .79 range" (p.32). Alderson *et al.* add that "if the test contains sections testing different skills in different ways, these sections will not correlate highly with one another, and the reliability will be lower" (p.89). Carroll and Hall (1985:127) suggest that, in judging reliability, a coefficient of above .75 is considered good but a preferable reliability coefficient should be about .90. Carroll and Hall (op. cit.) also suggest that for inter-judge (inter-rater) assessment, a reliability coefficient should be at least .60, and preferably between 0.70 and 0.80.

Thus, in order to obtain the reliability coefficient of the three sub-tests, Reading, Grammar, and Dictation, I calculated every item of the sub-tests using the SPSS programme. The Reading Test is discussed first, followed by the Grammar Test and then the Dictation Test.

5.3.1 The reliability analysis of the Reading Test

Table 5-7 below summarises the statistics for the reliability analysis of the Reading Test for samples from the AIS (N=413).

Table 5-7: Statistics for SCALE ALPHA for the Reading Test (N=413)

```
R E L I A B I L I T Y   A N A L Y S I S  -  S C A L E   (A L P H A)

                                                        N of
Statistics for        Mean    Variance     Std Dev     Variables
```

SCALE	31.0194	118.5142	10.8864	75

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
VAR00001	30.3196	114.3345	.4044	.9025
VAR00002	30.6126	112.6991	.5337	.9011
VAR00003	30.7627	116.5309	.1898	.9044
VAR00004	30.6320	115.4807	.2667	.9038
VAR00005	30.3293	114.2505	.4091	.9024
VAR00006	30.3002	114.6718	.3777	.9027
VAR00007	30.1332	117.0332	.2006	.9041
VAR00008	30.3123	114.0842	.4338	.9022
VAR00009	30.2712	116.3583	.2098	.9042
VAR00010	30.0944	117.1440	.2278	.9039
VAR00011	30.2082	116.7429	.1910	.9043
VAR00012	30.1186	117.5126	.1405	.9044
VAR00013	30.2857	115.8162	.2627	.9038
VAR00014	30.4019	115.1779	.2968	.9035
VAR00015	30.4116	116.1214	.2044	.9044
VAR00016	30.4116	113.8350	.4257	.9022
VAR00017	30.3632	114.0183	.4204	.9023
VAR00018	30.5787	114.8658	.3192	.9033
VAR00019	30.6053	116.5696	.1598	.9049
VAR00020	30.0847	117.5583	.1667	.9042
VAR00021	30.7772	117.1250	.1298	.9049
VAR00022	30.4479	117.6799	.0548	.9059
VAR00023	30.4237	115.7593	.2377	.9041
VAR00024	30.3390	115.7149	.2570	.9038
VAR00025	30.7167	117.2229	.1084	.9052
VAR00026	30.1162	116.6175	.2829	.9036
VAR00027	30.5981	116.2798	.1866	.9046
VAR00028	30.6271	117.0499	.1158	.9053
VAR00029	30.2373	118.1571	.0207	.9057
VAR00030	30.4673	113.2544	.4730	.9017
VAR00031	30.9370	116.0203	.4079	.9029
VAR00032	30.9225	116.4260	.3131	.9034
VAR00033	30.8935	116.3818	.2822	.9035
VAR00034	30.2179	115.9961	.2742	.9036
VAR00035	30.4455	113.3302	.4685	.9018
VAR00036	30.6005	113.6919	.4346	.9021
VAR00037	30.6416	114.2596	.3873	.9026
VAR00038	30.8644	115.6418	.3518	.9030
VAR00039	30.7627	113.3125	.5382	.9013
VAR00040	30.8378	115.9760	.2874	.9035
VAR00041	30.9734	117.6231	.1862	.9041
VAR00042	30.6102	114.6171	.3467	.9030
VAR00043	30.9903	118.2086	.0758	.9045
VAR00044	30.7651	113.3355	.5374	.9013
VAR00045	30.3850	114.0140	.4144	.9023
VAR00046	30.4237	113.9584	.4113	.9024
VAR00047	30.3390	114.0013	.4308	.9022
VAR00048	30.8668	114.3147	.5287	.9017
VAR00049	30.9709	117.8632	.1296	.9044
VAR00050	30.9613	116.8966	.3085	.9035
VAR00051	30.9661	117.1930	.2610	.9038
VAR00052	30.9709	117.2079	.2708	.9038
VAR00053	30.8184	116.2946	.2379	.9039

VAR00054	30.8789	115.4319	.3962	.9027
VAR00055	30.7918	116.0633	.2515	.9038
VAR00056	30.6465	114.8310	.3324	.9031
VAR00057	30.8305	116.0877	.2692	.9036
VAR00058	30.7337	112.9483	.5577	.9010
VAR00059	30.5448	111.2874	.6614	.8998
VAR00060	30.4504	112.5297	.5455	.9010
VAR00061	30.9322	116.6701	.2892	.9035
VAR00062	30.6102	112.2821	.5741	.9007
VAR00063	30.4262	112.4102	.5621	.9008
VAR00064	30.7094	114.7940	.3533	.9029
VAR00065	30.3269	116.0361	.2275	.9041
VAR00066	30.9976	118.2015	.0917	.9045
VAR00067	30.9903	118.2281	.0705	.9046
VAR00068	30.9709	117.7322	.1578	.9043
VAR00069	30.4407	113.6451	.4388	.9021
VAR00070	30.8208	115.3950	.3449	.9030
VAR00071	30.8814	115.3864	.4055	.9027
VAR00072	30.9031	116.8984	.2180	.9040
VAR00073	30.6247	113.3709	.4706	.9018
VAR00074	30.7942	114.3435	.4468	.9021
VAR00075	30.9758	117.4897	.2218	.9040

Reliability Coefficients

N of Cases = 413.0 N of Items = 75 Alpha = **.9044**

To analyse the data in Table 5-7 above, I use the guide suggested by Green and Weir (1998):

(a) The statistics for SCALE at the top of Table 5-7 supply us with similar information to that which we obtained when we described the descriptive analysis of the test (see descriptive analysis in 5.2.3.1 above).

(b) The first two columns of the Item-total Statistics tell us what would happen to the scale statistics if the question were to be removed. For instance, if Question 1 in the Reading Test (VAR00001) (see Appendix A.2.3: 470) were to be discarded, the total mean would drop from its current 31.0194 to 30.3196. The same applies to the scale variance: the total variance would drop from 118.5142 to 114.3345 if Question 1 were to be eliminated. This means that the removal of Question 1 (VAR00001) would mean a drop of 0.6998 in the scale mean and 4 points in the

scale variance. Since the item's discriminating power is observed for every question in the test, we could therefore see that some questions contribute more to the variance than others. This indicates that the bigger the discriminating power the questions have, the more they contribute to the scale variance. For example, for Question 59 (VAR00059) (see Appendix A.2.3: 474), the biggest contributor to the scale variance, we note that the scale variance would be 111.2874 if this item were to be dropped. This means that the scale variance drops 7.2295 points. An investigation into the item discrimination (ID) analysis (see Appendix B.2: 540) reveals that this question has an ID index of .88. In contrast, if Question 67 (VAR00067) (see Appendix A.2.3: 474), the smallest contributor to the scale variance, were to be dropped, the scale variance would drop from 118.5142 to 118.2281, less 0.2861 only. An investigation into the item discrimination (ID) analysis reveals that this question has ID index of .02 only. In other words, this question did not discriminate well and omitting or retaining it in the test would not have a big influence in discriminating between the lower and upper groups.

(c) The third column provides the information about each item's corrected item-total correlation (CITC), i.e. "...the correlation of the item with the total minus that item" (Green & Weir, 1998:50). Green & Weir (op. cit:50) add that: "given that the total test score is made up of all the reading test items, we would expect there to be a positive relationship between [total score of the Reading Test] and each individual item... Where the relationship is weak, we might suspect that the item concerned is varying in a different way" [not in the component of the Reading Test items]. If we look at the CITC in the third column above, we find that all questions have a correlation of above 0 (0 means no correlation at all). Some

questions have a scale of .1 or less which indicates a very low correlation. These questions, according to Green and Weir (op. cit.) demand further attention. The CITC is also an index of how a question discriminates between those candidates who are performing well on the overall test, and those who are performing badly (Green & Weir, op. cit.). This means that the more this is the case, the higher the discrimination. For example, if we look back at Question 59 (VAR00059) above, we find that the IF for this question was only .48, in other words, only about 50% of the candidates got this question correct . However, when we look at the CITC for this question , we find that it is .6614, the highest question correlates with the total minus of that item. At the same time, as described earlier in (b), this is the question that has the highest ID index.

(d) The final column, the Alpha If Item Deleted (AIID), "...tells us whether the test's internal reliability (here referred to Alpha) would increase or decrease if the particular item were removed" (Green & Weir, op. cit.). The way of interpreting the AIID is almost the same as the first two columns. The smaller the number of the AIID, the greater the item's contribution to the test overall Alpha. For example, the test overall Alpha, as shown at the end of Table, is .9044. The smallest AIID in column four is .8998, i.e. for Question 59 (VAR00059). This means that if this question were to be eliminated, the test overall Alpha would decrease by .0046. However, it seems that if some questions were to be removed, this removal would not have any effect on the test overall Alpha while, if others were to be eliminated, it seems that the test overall Alpha would increase. Examples of the former case are Questions 3 (VAR00003), 12 (VAR00012), 15 (VAR00015), 49(VAR00049) (see Appendix A.2.3: 470, 472, 474), and examples

of the latter are Questions 19, (VAR00019), 20 (VAR00020), 21 (VAR00021), 22 (VAR00022), 25 (VAR00025), 27 (VAR00027), 28 (VAR00028), and 29 (VAR00029) (see Appendix A.2.3: 473-74). If we look back at the IF and ID indices for the Reading Test (see Appendices B.1: 537 and B.2: 540), we may conclude, based on the comparison between the AIID and the IF and ID indices, that any questions that have high IF or ID indices or both will contribute to the increase in the test overall Alpha. On the contrary, any questions that have low IF or ID indices or both will not contribute to the increase in the test overall Alpha.

(e) The Reliability Coefficients at the bottom of Table 5-7 give us the summary of the total cases (candidates) who took the test, the number of items or questions in the test, and the Reliability Coefficient (r_{xx})of the test's consistency. The most important element of this description is the reliability coefficient because it shows us "...how consistent the test is in terms of the percentage of the reliable variation in scores, as opposed to error" (Brown,1988: 100). With regard to the final version of the Reading Test, the coefficient of r_{xx} =.9044 means 90% of the variation is reliable, i.e. related to variation in true score, and the remaining 10% is error, i.e. owing to chance. We may conclude that a coefficient of r_{xx} = .9044 indicates that the Reading Test is reliable. Since Lado (1961) and Alderson *et al.* (1996) suggest above that the vocabulary and reading tests are considered good if the coefficient of r_{xx} = .90, we may say that the final version of the Reading Test fulfills this requirement.

5.3.2 The reliability analysis of the Grammar Test

Table 5-8: Statistics for SCALE ALPHA for the Grammar Test (N=413)

R E L I A B I L I T Y A N A L Y S I S S C A L E (A L P H A)

Statistics for	Mean	Variance	Std Dev	N of Variables
SCALE	24.3680	82.3351	9.0739	50

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
VAR00001	23.7530	78.9243	.3666	.8843
VAR00002	23.6901	78.7678	.4032	.8838
VAR00003	23.6271	79.6228	.3219	.8849
VAR00004	24.1114	79.2982	.3654	.8843
VAR00005	24.0508	78.1115	.4865	.8826
VAR00006	23.6755	79.4430	.3252	.8849
VAR00007	23.7651	79.0539	.3491	.8845
VAR00008	23.9007	80.0314	.2298	.8863
VAR00009	23.9322	78.6944	.3853	.8840
VAR00010	23.8160	78.2330	.4376	.8832
VAR00011	23.7748	78.4953	.4128	.8836
VAR00012	23.8039	79.9590	.2399	.8862
VAR00013	24.2107	83.1036	-.1356	.8899
VAR00014	24.0169	80.7691	.1557	.8873
VAR00015	23.7264	77.5293	.5412	.8817
VAR00016	23.8039	76.8862	.5975	.8808
VAR00017	23.6683	79.2756	.3486	.8845
VAR00018	24.1283	79.3160	.3726	.8843
VAR00019	24.1646	79.5456	.3654	.8844
VAR00020	23.8015	78.4605	.4128	.8836
VAR00021	23.9056	79.9886	.2349	.8863
VAR00022	23.5593	79.6500	.3599	.8845
VAR00023	23.8232	79.6896	.2693	.8857
VAR00024	24.0121	77.8081	.5081	.8822
VAR00025	23.5884	79.5729	.3498	.8846
VAR00026	24.1743	80.6734	.2118	.8862
VAR00027	23.7700	79.2406	.3265	.8849
VAR00028	24.0702	79.6965	.2971	.8853
VAR00029	23.8668	79.3633	.3051	.8852
VAR00030	24.1090	80.0779	.2630	.8857
VAR00031	24.1065	79.7508	.3042	.8851
VAR00032	23.7869	77.9933	.4697	.8827
VAR00033	24.0436	80.1729	.2314	.8862
VAR00034	24.1041	79.8702	.2878	.8854
VAR00035	23.7240	79.0304	.3608	.8844
VAR00036	23.9249	80.1667	.2157	.8865
VAR00037	23.8281	78.2446	.4351	.8833
VAR00038	24.0872	79.3322	.3494	.8845
VAR00039	23.8257	79.2802	.3159	.8850
VAR00040	24.0097	78.3106	.4465	.8831
VAR00041	24.2058	79.8580	.3548	.8846

VAR00042	24.1308	80.2499	.2495	.8858
VAR00043	23.8499	78.5065	.4036	.8837
VAR00044	23.9274	77.8200	.4867	.8825
VAR00045	23.8789	77.9513	.4677	.8828
VAR00046	23.6465	79.3407	.3493	.8845
VAR00047	23.7167	78.2569	.4561	.8830
VAR00048	23.5303	80.4584	.2628	.8856
VAR00049	23.5787	80.7250	.1968	.8864
VAR00050	23.8281	78.2786	.4312	.8833

Reliability Coefficients

N of Cases = 413.0 N of Items = 50 Alpha = **.8866**

(a) The statistics for SCALE at the top of Table 5-8 supply us with similar information to that which we obtained when we described the descriptive analysis of the test (see descriptive analysis in 5.2.3.2 above).

(b) The first two columns of the Item-total Statistics tell us what would happen to the scale statistics if the item were to be discarded. A comparison with the item discrimination (ID) indices (see Appendix B.2: 541) shows that, in general, any questions that had an ID index of .40 and above would make the scale mean in the first column drop between 23.500 and 23.900 if these questions were to be deleted. On the contrary, any questions that had an ID index of .39 and below would make the scale mean in the same column drop from 23.900 to 24.200. The same happens to the scale variance in column two: for any items that have an ID index of .40 and above, the scale variance drops between 76.00 and 79.000. For example, if Question 16 which has an ID index of .85 were to be removed, the scale variance would drop from its current 82.3351 to 76.8862. On the other hand, with any questions that have an ID index below .39, the scale variance drops between 80.000 to 82.000 only. We may say here that, based on the data analysis in the first two columns, the questions with high ID indices contribute more to the mean and variance than others.

(c) If we look at the CITC in the third column above, we find that all except one question have a correlation of above 0 (0 means no correlation at all). Three questions have a scale of .1 or less which indicates a very low correlation, and, these need to be given further attention. With regard to Question 13 (VAR00013) (see Appendix A.2.3: 476), the CITC of this question is -.1356 which indicates that the question varies in a different way. An investigation into the ID analysis reveals that the ID index for this question was -.12 which means it did not discriminate at all. This was unexpected because the ID index of this question in the pilot study for both samples from Jordan and Malaysia was .42 and .18 respectively.

(d) In the final column, the lowest AIID in column four is .8808 (VAR00016). This means that if this question were to be eliminated, the test overall Alpha would decrease .0058. This supports our observation above that questions with high ID indices will increase the test's overall Alpha. In the meantime, if we want to increase the consistency of the test, any questions that have an AIID higher than the test's overall Alpha should be removed. Examples of these questions are Questions 13 (VAR00013) and 14 (VAR00014) (see Appendix A.2.3: 476).

(e) The Reliability Coefficients at the bottom of Table 5-8 tell us that the coefficient of r_{xx} for the Grammar Test is .8866, which means 89% of the variation is reliable, i.e. related to variation in true score, and the remaining 11% is error, i.e. owing to chance. We thus conclude that a coefficient of r_{xx} = .8866 indicates that the Grammar Test is reliable. Since Lado (1961) and Alderson *et al.* (1996) have suggested above that the structure test is considered good if the coefficient of r_{xx} = .90, we may conclude that the final version of the Grammar Test has almost fulfilled this requirement.

5.3.3 The reliability analysis of the Dictation Test

Table 5-9: Statistics for SCALE ALPHA for the Dictation Test (N=413)

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

Statistics for	Mean	Variance	Std Dev	N of Variables
SCALE	13.7215	32.6917	5.7177	25

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
VAR00001	13.0944	30.3381	.3973	.8732
VAR00002	13.0412	31.2095	.2423	.8774
VAR00003	12.9225	30.8241	.3831	.8734
VAR00004	13.1792	29.0358	.6338	.8661
VAR00005	12.8814	30.6437	.4711	.8715
VAR00006	13.0751	29.8900	.4916	.8705
VAR00007	12.9080	30.7051	.4245	.8725
VAR00008	13.1913	29.5580	.5308	.8693
VAR00009	13.2203	29.2790	.5837	.8677
VAR00010	13.2179	29.2728	.5849	.8676
VAR00011	13.2857	29.4958	.5469	.8688
VAR00012	13.4794	30.3424	.4583	.8715
VAR00013	12.9153	30.5146	.4622	.8716
VAR00014	13.4019	29.8866	.5068	.8701
VAR00015	13.1671	29.4647	.5515	.8687
VAR00016	13.3898	30.9617	.2874	.8763
VAR00017	13.0436	29.8767	.5076	.8701
VAR00018	13.2930	30.6979	.3184	.8756
VAR00019	12.8596	31.8395	.1881	.8775
VAR00020	13.0339	29.8823	.5113	.8700
VAR00021	13.3535	29.2242	.6195	.8667
VAR00022	13.4964	31.5613	.2033	.8779
VAR00023	13.0412	31.2823	.2281	.8778
VAR00024	13.6029	31.3177	.3503	.8742
VAR00025	13.2228	29.9891	.4472	.8718

Reliability Coefficients

N of Cases = 413.0 N of Items = 25 Alpha = **.8765**

(a) The statistics for SCALE at the top of Table 5-9 supply us with similar information to what we obtained when we described the descriptive analysis of the test (see descriptive analysis of the Dictation Test in 5.2.3.4 above).

(b) With reference to the first two columns, a comparison with the item discrimination (ID) indices (see Appendix B.2: 542) indicates that the relationship between the scale mean if item deleted and the ID indices of the test is not strong. This means if questions with high ID indices were to be eliminated, the scale mean would not drop in a greater point compared to questions with low ID indices. For example, Question 4 (see Appendix A.2.5: 505) has an ID index of .86, i.e. discriminates very well. If this question were to be removed, the scale mean would drop from its current 13.7215 to 13.1792 only. On the contrary, Question 19 (see Appendix A.2.5: 505) has an ID index of .23, much lower than the ID index for Question 4 above. If this question were to be eliminated, the scale mean would be dropped to 12.8596, much more than the scale mean if Question 4 were to be eliminated. With reference to the second column, the scale variance, it is noted that the degree the total number of the scale variance would drop depends on the ID indices of the question: if the questions have high ID indices, the scale variance would drop more; if the questions have low ID indices, the scale variance would drop less. For example, if Question 4, which had an ID index of .86, were to be removed, the scale variance would drop from its current 32.6917 to 29.0358, i.e. less 3.6559 points. On the contrary, if Question 19 , which had an ID index of .23, were to be removed, the scale variance would drop to 31.8395 only, i.e. less than one point. We may conclude here that, based on the distribution of the analysis data in the scale variance, questions in the Dictation Test have a discriminating power because questions with high ID indices contribute more to the variance than others.

(c) If we look at the CITC in the third column above, we find that all except one question have a correlation of above .1 (.1 means the correlation is very low). 16

questions have correlation for an item with the total minus that item, ranging between .4 and .6. Question 19 has a scale of .1881 which indicates a very low correlation and needs to be given further attention.

(d) In the final column, the interpretation of the AIID is the same as the first two tests above. This means that if questions that have high ID indices were to be eliminated, the test overall Alpha would drop more compared to questions that have low ID indices. This supports the conclusion we came to earlier that questions with high ID indices will increase the test's overall Alpha. In the meantime, if we want to increase the consistency of the test, any questions in the last column that have an AIID higher than the test's overall Alpha would have to be removed.

(e) The Reliability Coefficients at the bottom of Table 5-9 tell us that the coefficient of r_{xx} for the Grammar Test is .8765, which means 88% of the variation is reliable, i.e. related to variation in true score, and the remaining 12% is error, i.e. owing to chance. We thus conclude that a coefficient of r_{xx} = .8765 indicates that the Dictation Test is reliable. Since Lado (1961) and Alderson *et al.* (1995) suggest that the auditory comprehension tests are considered good if the coefficient of r_{xx} = .80-.89, we may say that the final version of the Dictation Test has achieved this requirement.

5.3.4 The reliability analysis of the Essay Test

As stated earlier, the Reliability coefficient could not be obtained through internal consistency analysis for the Essay Test because all items were unequally weighted and were not marked as either right or wrong. Instead some aspects of the

assessment have a total mark of ten while others have a total mark of five. There are several ways to monitor the degree of consistency for the oral and essay tests. Two terms appear often in the discussion: *intra-rater reliability* and *inter rater reliability* (Alderson *et al.* 1996:129). The former refers to the consistency of the markers in giving marks to the same scripts or oral performances on two different occasions while the latter refers to the degree of similarity between different examiners in giving marks to the same scripts or oral performances (Alderson.*et al.* op. cit.). Carroll and Hall (1985:121) suggest that "...a simple way to check on inter-marker [inter-rater] reliability is to use the ranking method". This method is implemented by giving a number of scripts to two examiners and then ask them to mark them independently according to the marking scheme. Then, using the following formula, the Rank Order Correlation (ROC) is calculated (the formula was taken from Carroll and Hall, op. cit:119):

$$R = 1 - \frac{6 \times \text{Total } d^2}{n(n^2 - 1)}$$

where: R = Rank-order correlation
Total d^2 = the total of differences squared between two examiners
n = the number of scripts marked

As stated in 5.2.2.1 (the administration of marking the subjective test), three examiners were asked to mark 10 scripts from their colleagues for the purpose of the Reliability analysis. I personally marked 10 scripts which brought the total to 40. These scripts were then divided into four groups for the inter-marker reliability. Tables 5-10 to 5-13 below show the details of marks for the four groups of examiners together with the total of differences squared between two examiners (d^2):

Table 5-10: Marks of the first group of the examiners

student id	rater 1	rater 2	d	d^2
1	13	15	2	4
2	8	8	0	0
3	13	15	2	4
4	9	11	2	4
6	7	7	0	0
7	9	10	1	1
8	11	12	1	1
9	17	17	0	0
10	7	7	0	0
12	12	14	2	4
			Total d^2	18

Table 5-11: Marks of the second group of the examiners

students id.	rater 1	rater 2	d	d^2
138	10	11	1	1
139	16	16	0	0
140	14	14	0	0
141	5	7	2	4
142	15	17	2	4
143	11	11	0	0
144	13	13	0	0
145	14	16	2	4
146	9	12	3	9
147	12	15	3	9
			Total d^2	31

Table 5-12: Marks of the third group of the examiners

students id.	rater 1	rater 2	d	d^2
259	24	24	0	0
260	12	10	2	4
261	9	10	1	1
262	13	15	2	4
263	17	14	3	9
265	16	16	0	0
266	14	11	3	9
267	20	19	1	1
268	19	19	0	0
270	15	13	2	4
			Total d^2	32

Table 5-13: Marks of the fourth group of the examiners

students id.	rater 1	rater 2	d	d^2
376	16	16	0	0
377	11	13	2	4
378	22	18	4	16
379	18	16	2	4
382	15	17	2	4
383	15	14	1	1
384	9	12	3	9
387	15	15	0	0
388	10	10	0	0
389	11	13	2	4
			Total d^2	42

I demonstrate below how to calculate the rank order correlation (ROC) between the raters (the calculation uses the first pair's marks as a sample)

$$R = 1 - \frac{6 \times 18}{10(100-1)}$$

$$R = 1 - \frac{108}{990}$$

$$R = 1 - 0.109 = 0.891$$

Thus the ROC for the first group of the examiners is 0.891. Using the same formula, I calculated the ROC for the remaining three groups. Below are the results:

The ROC for the second group, in Table 5-11, is 0.812, the third group, in Table 5-12, is 0.807, and the last group, in Table 5-13, is 0.746. This degree of correlation indicates a fairly high order of relationship between the judgement of the eight raters (Carroll and Hall, op. cit.). Green and Weir (1998:110) however view that if the correlation between the markers is around .7, some improvements need to be made in the way the examiners mark the scripts. This is because such a correlation indicates that the markers agreed in only 49% (.7 x .7) (Green and Weir, op. cit.).

5.4 The correlation coefficient analysis

The correlation coefficient (r) is a statistic which expresses the degree of relationship between two sets of test scores or variables (Harris, 1988:142). The correlation coefficient that I am using in this analysis is called the *Pearson product-moment correlation coefficient* which is the statistic of choice for comparing two sets of variables or data that use the interval scales. The purpose of calculating the coefficient is to determine how two different variables or scores of the candidates on a sub-test correlate with each other. From the calculation, we can describe whether or

not the correlation is strong and in the same direction.

5.4.1 Types of correlation

Rowntree (1991:160) divides correlation into three types, namely *positive*, *negative*, and *zero* correlations. Positive correlation refers to "...the changes in one variable [that] are accompanied by changes in the other variable and in the *same* direction [i.e.] the larger values on one variable tend to go with larger values on the other". This type of correlation can take the relationship up to the maximum value of $r = +1.0$ (Brown, 1996). Brown (op. cit.) adds that "...such a correlation [+1.0] occurs only if the two sets of data line up in exactly the same order .. and this is the reason such relationships are called *linear*" (p.153). Negative correlation refers to "...the changes between the two variables or values in *opposite* directions. Larger values on one will tend to go with smaller values on the other" (Rowntree, op. cit:160) and "...can be negative in value as high in magnitude as $r = -1.0$"(Brown, op. cit.). This indicates that if students score high on one test, they score low on the other, or vice versa. Zero correlation, $r = 0$, refers to "...no clear tendency for the values on one variable to move in a particular direction (up or down) with changes in the other variable" (Rowntree op. cit.). However, perfect zero correlation is unlikely to occur in statistical inquiries because "...even random numbers may haphazardly produce a correlation coefficient of some magnitude" (Brown op. cit:162).

5.4.2 Interpreting a correlation coefficient

Interestingly, it is difficult to determine at what counts the correlation coefficient can be considered weak or strong. Some statisticians, like Rowntree and others, are of the view that it depends on the samples: "...in a sample of 1,000, a

correlation of $\pm$ 0.08 would be significant at the 1% level" (Rowntree, op. cit:169). Brown (op. cit.) adds that if we calculate a correlation coefficient of 100 random numbers from different samples (100 in one column and 100 others in another column) and then plot the relationship, we may observe some relationship, say r = -.0442, because these numbers may "...haphazardly produce a correlation coefficient of some magnitude" (p.162). However, as a guide, several levels have been suggested to determine whether the correlation coefficient is weak, or low, or moderate, or strong. Rowntree (op. cit:170), for example, suggests the following levels:

0.0 to 0.2	very weak, negligible
0.2 to 0.4	weak, low
0.4 to 0.7	moderate
0.7 to 0.9	strong, high, marked
0.9 to 1.0	very strong, very high

Hughes, (1992:160) simplifies the issue, stating that "...items that show correlation of 0.3 or more are generally considered satisfactory and to be contributing well to the total test". Rowntree (op. cit.) disagrees with Hughes's view by saying that it all depends on the context. Rowntree goes further, stressing that "...in the majority of cases the question of 'satisfactoriness' is totally irrelevant" (p.170). Green & Weir, (1998:109) are of the view that "...we might expect correlations of between .4 and .7 on different parts of the same test simply because of an underlying ability which underpins all language behaviour".

Before investigating the correlation coefficient of the scores in the final version of the test, I first have first to set up the null hypothesis (H_0) and consequently alternative hypothesis (H_1). The null hypothesis for the four subjects (Reading, Grammar, etc.) of this test is as follows:

There is no relationship among the performance of the candidates on their scores in the sub-tests, thus we assume that the $r = 0.00$

The alternative hypothesis is read as follows:

There is relationship among the performance of the candidates on their scores in the sub-test, thus we anticipate that the $r = > 0.00$.

Six correlation coefficients and relationships will be calculated and plotted on the total scores of the sub-tests using the SPSS. They are between the scores of:

(a) the Reading and the Grammar Tests

(b) the Reading and the Essay Tests

(c) the Reading and the Dictation Tests

(d) the Grammar and the Essay Tests

(e) the Grammar and the Dictation Tests

(f) the Essay and the Dictation Tests

Before plotting the scores, I have to decide the degree (significance) to which I want to be sure of my results because I cannot obtain 100% certainty. Brown (op. cit.) suggests that in language testing, such decisions are traditionally set at 95% or 99%. I decided to set the level at 99%. This level ensures that only a 1% chance exists (less than .01 probability ($p < .01$). With this certainty level, I can be 99% sure that I am right in rejecting the notion (null hypothesis) that my correlation coefficient is due to chance alone. Table 5-14 below displays the correlation coefficient of the scores that have been calculated together with the degree (significance) at 2 tails format:

Table 5-14: Correlation coefficient of sub-tests (N=413)

Correlations

		TTLREAD	TTLGRAM	TTLESSAY	TOTALDIC
TTLREAD	Pearson Correlation	1.000	.746**	.687**	.748**
	Sig. (2-tailed)	.	.000	.000	.000
	N	413	413	413	413
TTLGRAM	Pearson Correlation	.746**	1.000	.701**	.708**
	Sig. (2-tailed)	.000	.	.000	.000
	N	413	413	413	413
TTLESSAY	Pearson Correlation	.687**	.701**	1.000	.711**
	Sig. (2-tailed)	.000	.000	.	.000
	N	413	413	413	413
TOTALDIC	Pearson Correlation	.748**	.708**	.711**	1.000
	Sig. (2-tailed)	.000	.000	.000	.
	N	413	413	413	413

**. Correlation is significant at the 0.01 level (2-tailed).

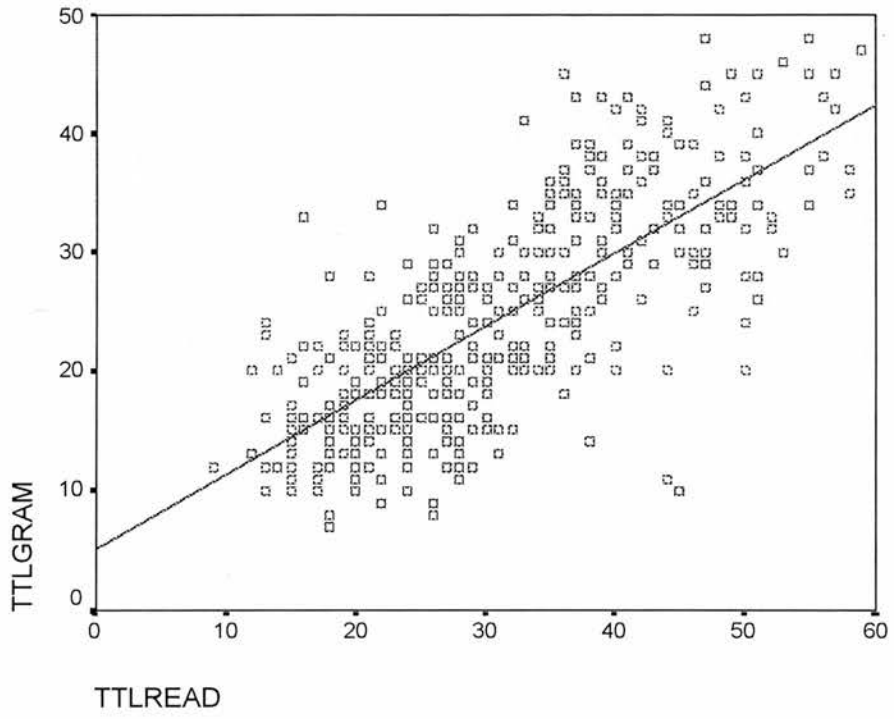
With reference to the first column in the left hand side, the correlation coefficient between TTLREAD and TTLGRAM is r = .75 (the two nearest decimal points) and is statistically significant at $p < .01$ (see figure .000 parallel to Sig. (2-tailed)). To put it another way, I can be 99% sure that the correlation coefficient occurred for reasons other than chance. For the correlation between the total scores of the Reading and Essay (TTLESSAY) tests, the r = .69 at $p < .01$ too. The correlation coefficient between the third pair, the Reading and the Dictation (TOTALDIC) tests is r = .75 at $p < .01$. With reference to the second column, the correlation coefficient between the Grammar and the Essay Test is r = .70 at $p < .01$. The second correlation that needs to be reported in the second column is between the total scores of the Grammar and the Dictation Tests: the r = .71. The last correlation that needs to be reported is between TTLESSAY and TOTALDIC in the third

column: the $r = .71$ at $p < .01$. From this result, we may stress here that the average correlation coefficients of the total scores for the sub-tests when these total scores are calculated are between $r = .70$ and .75. If we adopt Rowntree's suggestion (op. cit.), we may say that the correlation coefficients for the total scores of the tests are just enough to be described as strong, high and marked. If we accept the suggestion by Hughes (1992), we may say that the correlation coefficients of the test's scores are more than satisfactory and will contribute well to the total test (if calculated). If we consider the opinion of Green and Weir (1998), we may conclude that the *r*'s for these tests are within the 'acceptable' range of the correlation. With reference to the null hypothesis, since the probability of the occurrence by chance is always less than .01, I am obliged to reject H_0 because there is a correlation between the scores of different tests for the same candidates and this correlation does occur because of some factor other than chance.

To get a clearer picture and interpretation of the correlation and relationship for the candidates' scores, I use a simple *scatterplot*, "... a form of visual representation, similar to the histogram, bar graph, ... that allows for representing two sets of scores at the same time and examining their relationship" (Brown, op. cit:152). From the shape and slope of the plotted points, the correlation coefficient of two scores for the same candidates is considered strong if "the plotted points on the 'scatter' diagram lie on a straight line" (Rowntree, op. cit:161). This means that the nearer the plotted points to the straight line, the stronger the relationship between the two scores, and the more scattered the plotted points from the straight line, the weaker the relationship between the two scores. In addition, to see how significant was the correlation coefficient for the test (in the hypothetical distribution of

correlation coefficients), I calculated the Standard Error of the Correlation Coefficient (SE_r). According to Rowntree, (op. cit:166), "...we can estimate the SE_r by squaring the correlation coefficient, subtracting it from one, and then dividing it by the square root of the number of pairs in the sample". The description of the scatter plots below starts first with the correlation coefficient for the scores of the Reading and Grammar Tests:

Figure 5-5: Scatterplot for the Reading and Grammar Tests



From Figure 5-5, we note that the scores for the Reading Test (TTLREAD) were on the x axis (horizontal) while the scores for the Grammar Test (TTLGRAM) were on the y axis (vertical). The end line for the Reading Test is 60 because the highest mark for this test was 59 and the end line for the Grammar Test is 50 because the highest mark for this test was 48. It should be noted here that the position of the variables (in this case the scores) is not affected by their position, i.e. horizontal or vertical because the purpose of pooling these variables is to look for a relationship and not the effect of one variable on another. From the shape and slope of the squares, we can see clearly that the type of scatterplot for this correlation is called a positive one, i.e. larger values on the Grammar Test scores tend to go with larger values on the Reading Test and vice versa. However, a small number of scores are scattered

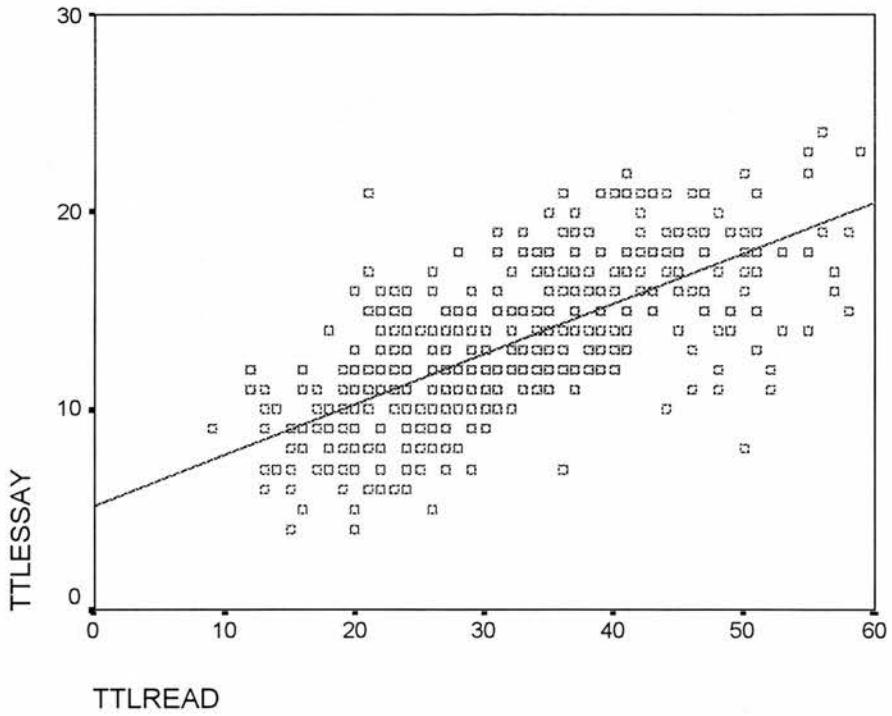
indicating the likelihood of some differences between the variables. This is justified because the Reading Test was considered more difficult than the Grammar Test. The SE_r for this correlation, using the above calculation, can be calculated as 0.02. This means that if we were to conduct the correlation coefficient of the same test, we would expect to find that about 68% of the samples had correlation coefficients in the range $r \pm 1SEr$, i.e. 0.75 ± 0.02. For about 95% of the samples the correlation coefficients would be between $r \pm 2SEr$, i.e. 0.75 ± 0.04. And finally, about 99.7% of the samples would lie between $r \pm 3SEr$, i.e. 0.75 ± 0.06. To summarise, if the correlation coefficients of the scores for the Reading and Grammar Tests are calculated in a hypothetical distribution of r using the same number of candidates, we would be sure the r for these scores is as follows:

68% sure r lies between + 0.73 and +0.77

95% sure r lies between +0.71 and +0.79

99.7% sure r lies between +0.69 and 0.81

Figure 5-6: Scatterplot for the Reading and Essay Tests



The interpretation of the scatterplot for the correlation between the scores of the Reading Test and the scores of the Essay Test is as described above: there is some evidence of a positive relationship between the two variables (scores): as the students' scores on TTLREAD increase so do their scores on TTLESSAY. It is observed however that the relationship between the two scores was closer at the bottom and started to scatter when they reached the top. It is also noted that a small number of those who obtained good scores in the Reading Test obtained lower marks in the Essay Test (small squares plotted under the straight line). This could be one of the reasons why the correlation coefficient between these two scores was the lowest in the rank: less than .70. The SE_r, for these scores is 0.03. If we total up the

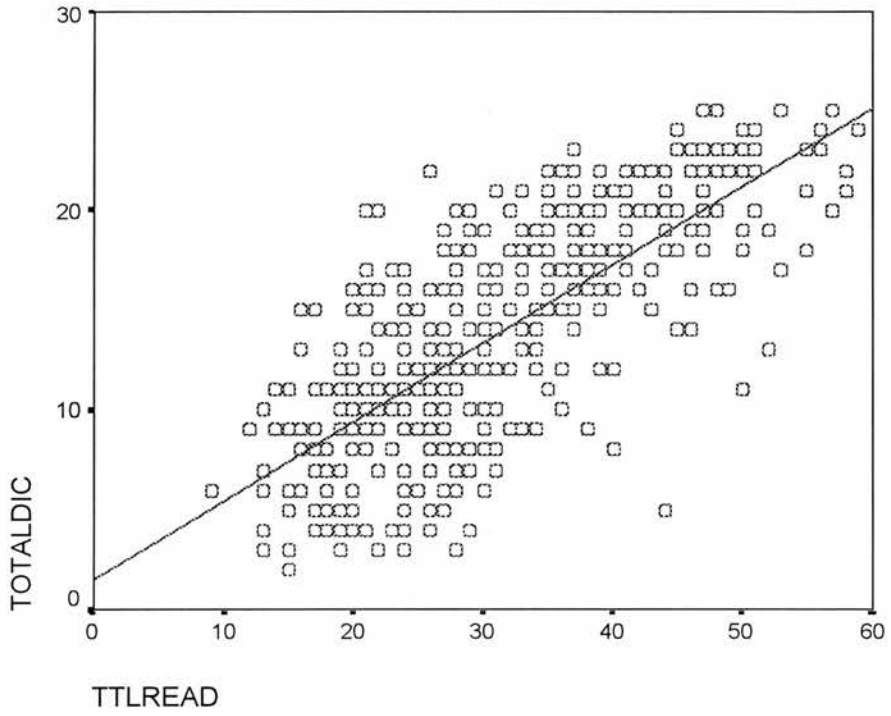
correlation coefficients of the scores for the Reading and the Essay Tests using the SE_r, the sure calculation of r would be:

68% sure r lies between + 0.66 and +0.72

95% sure r lies between +0.63 and +0.75

99.7% sure r lies between +0.60 and 0.78

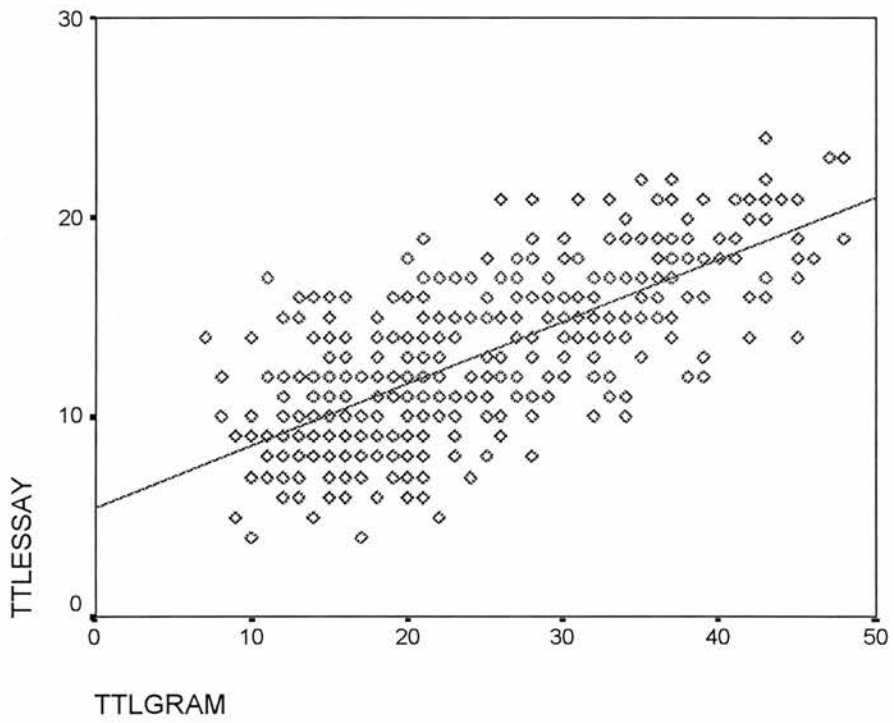
Figure 5-7: Scatterplot for the Reading and Dictation Tests



From the shape and slope of the quadrangles, we note that there is evidence of a positive relationship between the scores of the Reading and the scores of the Dictation Test. This boils down to saying that as the candidates' scores on TTLREAD increase so do their scores on TOTALDIC. However, more than half of the population of those who scored between 20 and 30 on TTLREAD secured very low marks in TOTALDIC, ranging from 2 to 5 only. On the contrary, those who obtained marks between 15 and 25 on TOTALDIC seemed to secure between 30 and

45 or 50 on TTLREAD. In my view, this phenomenon created a balance for the relationship between these two scores and therefore resulted in a quite strong correlation coefficient: r = .75. The SE_r for these scores is the same with the SEr between TTLREAD and TTLGRAM, i.e. 0.02, because of the same r. Hence, if I totaled up the correlation coefficients of the scores for both tests using the SE_r, the calculation would be exactly as the first pair above (see scatterplot for the Reading and Grammar Test).

Figure 5-8: Scatterplot for the Grammar and Essay Tests



From the shape and slope of the diamonds in Figure 5-8, we can see that there is clear evidence of a positive relationship between the two scores, that is, as the candidates' scores on TTLGRAM increase so do their scores on TTLESSAY. However, the plotted points are more scattered when both scores reach higher marks.

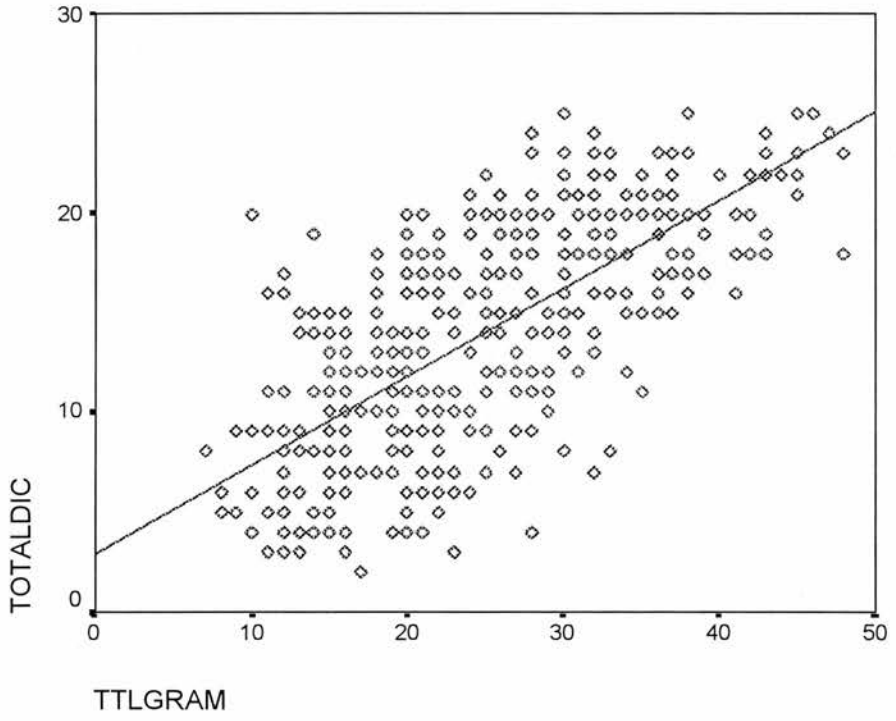
It is also noted that the plotted points (candidates' scores) of both tests are concentrated more at the bottom and middle parts rather than the upper part, i.e. more candidates scored lower and average marks than higher marks. The SE_r, for these scores is 0.03. If I totaled up the correlation coefficients of the scores for the Reading and the Essay Tests using the SE_r, the sure calculation of *r* would be:

68% sure *r* lies between + 0.67 and +0.73

95% sure *r* lies between +0.64 and +0.76

99.7% sure *r* lies between +0.61 and 0.79

Figure 5-9: Scatterplot for the Grammar and Dictation Tests



From the shape and slope of the diamonds in Figure 5-9, we can see that there is some evidence of a positive relationship between the two scores, that is, changes in the candidates' scores on TTLGRAM are accompanied by changes in the candidates' scores on TOTALDIC and in the same direction, i.e. towards the positive direction.

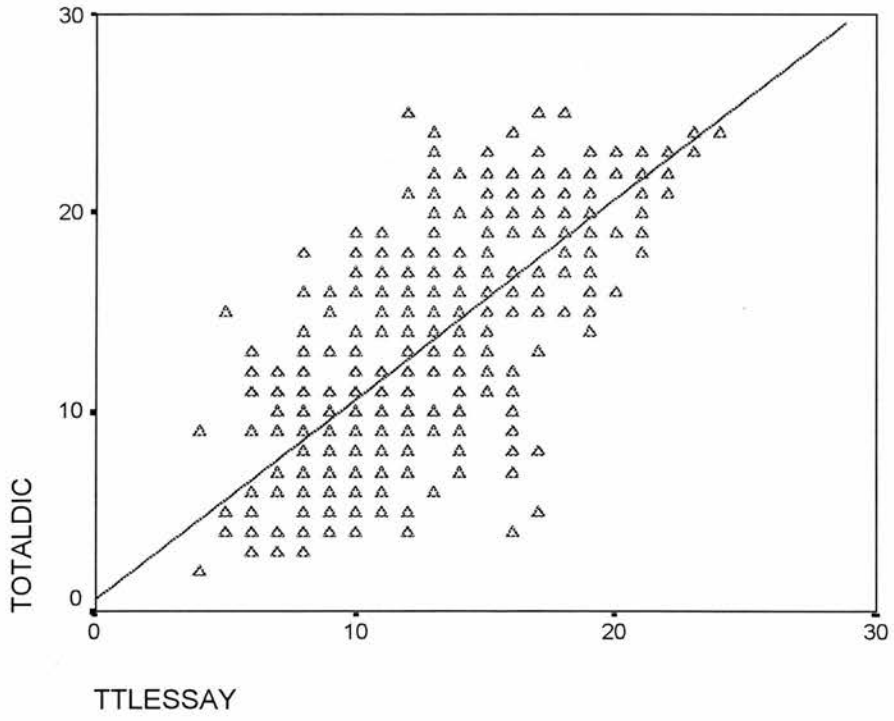
However, the plotted points for both scores are more scattered from the straight line if we compare this Figure with the previous Figures. A small number of candidates who scored average marks for TTLGRAM seemed to obtain low marks for TOTALDIC. This phenomenon did not affect the correlation coefficient of both scores very much; hence we observed the $r = 0.71$. The SE_r for these scores is 0.02. The total of the sure calculation of r of the scores for the Reading and the Essay Tests using the SE_r would be:

68% sure r lies between + 0.69 and +0.73

95% sure r lies between +0.67 and +0.75

99.7% sure r lies between +0.65 and 0.77

Figure 5-10: Scatterplot for the Essay and Dictation Tests



From the shape and slope of the triangles in Figure 5-10, we can see that the

relationship between two scores, that is, TTLESSAY and TOTALDIC are the same, i.e. towards the positive direction. The interpretation of the plotted points for both scores is almost the same as in the previous Figures, especially Figures 5-8 and 5-9. This is justified because these three pairs have very similar correlation coefficients: .70 and .71. The calculation of the SE_r for these pairs shows that they have the SE_r of 0.03 and the sure calculation of *r* would be:

68% sure *r* lies between + 0.68 and +0.74

95% sure *r* lies between +0.65 and +0.77

99.7% sure *r* lies between +0.62 and 0.80

It is noted from the above discussion that the true correlation of all pairs in the above Figures lie down within 3 standard errors. The reason is very simple: "...the bigger the correlation coefficient, the larger would be the number subtracted from 1, and the smaller the number into which the square root of the sample size is divided; and so the smaller the size of the standard error" (Rowntree op. cit:166). Moreover, the number of the samples acts positively in reducing the SE_r. "...The larger the sample, the larger the square root we divide by, and hence the smaller the standard error [of the correlation coefficient]" (op. cit.).

To draw the conclusion of the last aspect of the statistical analysis of the test, the correlation, I may stress here that two factors, the high coefficient and the relatively large number of samples, have contributed greatly to the significance of the relationship of total scores of these sub-tests. In addition, the proper selection of the items for the test, discussed earlier in the discussion of content validity, and the modification of these items in the course of the pilot study, have indeed contributed to the high correlation coefficient.

5.5 Concurrent validity

As discussed in Chapter Two (see 2.2.2.2.1), concurrent validity involves the comparison of the test scores with some other measures for the same candidates taken at roughly the same time as the test. Alderson *et al.* (1995) suggest that *'some other measures'* can be scores from a parallel version of the same test, the testees' self-assessments of their language abilities, or candidates' ratings on a number of relevant dimensions by teachers, subject experts or other informants. Gronlund (1982), Harris (1988) and Brown (1996) stress that the parallel test must be a test that is already a well-established measure of the construct involved. Brown (op. cit.) for example suggests that in the light of English language testing, the comparison should be with such tests as the *Test of English as a Foreign Language* (TOEFL) because this kind of test is a well-established one.

Some researchers think that it is difficult to obtain concurrent validity. Morrow (1979) (in Kattan 1990), for example, argues that, in English language testing, we cannot externally validate the test because there are no similar tests to compare it with. Clapham and Hughes (1988), (also in Kattan op. cit.), however, argue that although such validation studies are difficult, this does not mean this kind of validity is unnecessary. Clapham and Hughes (in Kattan, op. cit:275) add that "if we do not check the test against external criteria, how can we know whether it is assessing the candidates with any degree of accuracy?" Davies (1984) (in Kattan, op. cit.) shares this view with Clapham and Hughes when he insists that "...the external criterion, however hard to find and however difficult to operationalise and quantify, remains the best evidence of a test's validity" (p.275). Another difficulty

encountered, with reference to concurrent validation, is that we cannot claim that a test has validity simply because of a high correlation between any two tests, if the other test does not measure the same construct. Thus Hawkey (1982) (in Kattan op. cit.) states that "...other tests available as criteria for concurrent validation are likely to be less integrative/communicative in construct and format and thus not valid as references for direct comparison" (p.275).

After considering the above opinions, I discovered that it is indeed very difficult to obtain scores from the same students from a parallel test taken at roughly the same time as the placement test, especially if we search for a test at the level of TOEFL and the like in Arabic. I therefore decided to employ two types of measures: first the students' self-assessments of their language abilities; and second the results from the entrance examination which was not taken at the same time as the placement test, but which tested the same construct. The details of these two measures are discussed under the individual topics below. It is also important to stress here that I am not over optimistic about the results of these two measures. This is due to the fact that items in the parallel test, in this case the entrance examination, were not, to the best of my knowledge, statistically and empirically analysed. Therefore, if the correlation coefficient is found to be low or negative, this may be as a result of low content validity of the items in the parallel test. It may also be because the items in the parallel test, statistically, had a very low item facility or item discrimination.

5.5.1 The students' self-assessment

The effort to involve the learner in the concurrent validation is considered by some educationists as a new development in educational psychology which

emphasises the central role of learners. Rea (1985) for example, states that in the communicative curriculum, the learner is seen as an active participant at all stages of the teaching and learning process, including evaluation. Self-assessment, for example, aims to involve the learner in making evaluative judgements on his or her own performance. Alderson *et al.* (1995:177) also state that the concurrent validity of the test may be obtained by comparing the results of the candidates with "...the candidates' self-assessment of their language abilities...". Alderson *et al.* (op. cit.) further suggest that the comparison is "...usually expressed as a correlation coefficient ranging in value from -1.0 to +1.0". Alderson *et al.* however do not seem to be very optimistic as to whether the comparison of the results of the test with the self-assessment will produce a high correlation coefficient:

> "Most concurrent validity coefficients range from +.5 to +.7 - higher coofficients are possible for closely related and reliable tests, but unlikely for measures like self-assessments..."(p.178).

To obtain the student's self assessment of their ability in such a placement test, I prepared a simple questionnaire which included the specific skill areas and enabling skills that made up the design test (see Appendix A.3.2: 532-34). The questionnaire was divided into four parts: Reading, Grammar, Dictation, and Essay Tests. The students were asked to assess their ability using a four-point scale:

1 = very weak
2 = weak
3 = good
4 = very good

For the Reading Test, the questions included most of the reading skills that are incorporated in the test (see Appendix A.3.2: 532). There were nine questions: the

first six were aimed at the test while the last three were explicitly aimed at the test itself. With regard to the Grammar Test, five questions were asked (see Appendix A.3.2: 533). Question one included five sub-questions that were related to syntax while question two included three sub-questions that were related to morphology, which made a total of eleven questions altogether. Eight questions could be said to be implicit and the last three were explicitly related to the test itself. As for the Dictation Test, there were eight questions: all questions were explicitly related to the test (see Appendix A.3.2: 533-34). For the Essay Test, five questions were set up: each of them was related to the test through the marking scheme, i.e. the content and the organisation of ideas, the uses of grammar, vocabulary, and punctuation (see Appendix A.3.2: 534).

To prevent the students from being influenced by the test, I distributed the questionnaire two months after the placement test had taken place. Among the problems I encountered during the administration of this questionnaire was the difficulty of distributing the forms to the target students. Many of them were absent from lectures, which was the primary way to see them. Three weeks before the end of the session, 47 students, out of 413, had not yet answered the questionnaire. I had no other choice than to put their names on the notice board asking them to collect the forms from the AIS administration office, fill them in, and then return them to the same office.

Before analysing the result of the correlation coefficient, I establish below the null and alternative hypotheses for concurrent validity:

There is no correlation between the items of the test and the students' self-assessment. Therefore, the $r = 0.00$. (H_0)

There is a correlation between the items of the test and the students' self-assessment. (H_1)

5.5.1.1 The results

When analysing the result, the means for the students' self-assessment are displayed first, together with the means for the actual performance in the placement test for the purpose of comparative study. Then the results from the former will be plotted with the results from the latter to obtain the correlation coefficient (r) from which the confirmation as to whether the correlation between these two variables is calculated as high or low is derived. The analysis of the Reading Test is discussed first. This is then followed by the Grammar, the Dictation, and lastly the Essay Tests.

5.5.1.1.1 The Reading Test

After close scrutiny, I found that five criteria (see Table 5-15 below) from the questionnaire form were closely related to the test. With regard to the questions in the Reading Test, I chose several questions to be correlated with the criteria: nineteen (1-6, 13-15, 16-25) (see Appendix A.2.3: 470-73) for the first criterion, eleven (7-10, 11-12, 26-30) (see Appendix A.2.3: 471, 472, 474) for the second, the total marks for Part One for the third, the total marks for Part Two for the fourth, and the total marks for Part Three for the fifth criterion. Table 5-15 below displays the means for both the questionnaire and the Reading Test. Since the total marks for both variables differ from each other, the means are calculated in percentage.

Table 5-15: The means (in percentage) for the students' self-assessment and performance

Criteria:	self-assessment (%)	performance (%)
1. Understanding texts not related to the area of study	54	56
2. Understanding texts related to the area of study	69	73
3. Answering multiple-choice questions format	69	64
4. Answering true-false questions format	69	61
5. Answering cloze test format	54	28

We can see from Table 5-15 the means for the students' self-assessment and performance. The percentages for performance are obtained by calculating the means for all the questions related to the criteria, dividing these means by the total number of the questions and then multiplying by one hundred. From the data in Table 5-15, we may suggest the following: the means for the students' self-assessment for three criteria, 1, 2 and 3, are closely related to their actual performance. In other words, the overall opinion of the students' regarding their ability in these three criteria matched their actual performance in the test. However, the means for criterion 5 did not agree with the students' performance. The mean shows that the students overestimated their actual capability in the cloze test. For the fourth criterion, we may say that the means between self-assessment and performance are at the average level.

The percentages for both variables (self-assessment and performance) in Table

5-15 however, do not indicate the actual relationship of self-assessment and performance for the same candidate. To see the relationship between these two variables, we need to run the correlational analysis. Table 5-16 below displays the correlation coefficient (r) between the students' self-assessment and their performance for the first criterion:

Table 5-16: The *r* for understanding texts not related to the area of study

		READ5
VAR00001	**Pearson Correlation**	**-.099**
	Sig. (2-tailed)	**.044**
	N	**413**
VAR00002	**Pearson Correlation**	**-.012**
	Sig. (2-tailed)	**.804**
	N	**413**
VAR00003	**Pearson Correlation**	**-.004**
	Sig. (2-tailed)	**.939**
	N	**413**
VAR00004	**Pearson Correlation**	**-.047**
	Sig. (2-tailed)	**.339**
	N	**413**
VAR00005	**Pearson Correlation**	**.016**
	Sig. (2-tailed)	**.742**
	N	**413**
VAR00006	**Pearson Correlation**	**.041**
	Sig. (2-tailed)	**.405**
	N	**413**
VAR00013	**Pearson Correlation**	**.025**
	Sig. (2-tailed)	**.606**
	N	**413**
VAR00014	**Pearson Correlation**	**.045**
	Sig. (2-tailed)	**.364**
	N	**413**
VAR00015	**Pearson Correlation**	**.005**
	Sig. (2-tailed)	**.919**
	N	**413**
VAR00016	**Pearson Correlation**	**-.010**
	Sig. (2-tailed)	**.842**
	N	**413**
VAR00017	**Pearson Correlation**	**-.073**
	Sig. (2-tailed)	**.137**
	N	**413**
VAR00018	**Pearson Correlation**	**.000**
	Sig. (2-tailed)	**.996**
	N	**413**
VAR00019	**Pearson Correlation**	**-.016**
	Sig. (2-tailed)	**.748**
	N	**413**
VAR00020	**Pearson Correlation**	**.065**
	Sig. (2-tailed)	**.187**
	N	**413**
VAR00021	**Pearson Correlation**	**-.029**
	Sig. (2-tailed)	**.553**
	N	**413**
VAR00022	**Pearson Correlation**	**.074**
	Sig. (2-tailed)	**.132**
	N	**413**
VAR00023	**Pearson Correlation**	**.070**
	Sig. (2-tailed)	**.155**
	N	**413**
VAR00024	**Pearson Correlation**	**-.025**
	Sig. (2-tailed)	**.607**
	N	**413**
VAR00025	**Pearson Correlation**	**.035**
	Sig. (2-tailed)	**.480**
	N	**413**

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

From Table 5-16, we note that the correlation coefficients between the students' self-assessment of their ability (READ5) and their performance (VAR00001-VAR00025) (see Appendix A.2.3: 470-73) on understanding texts not related to the area of their study, are very low and not statistically significant at either $p < .01$ or $p < .05$. For example, the correlation coefficient between READ5 and VAR00018 has the $r = .00$. This means that no relationship exists between the two

sets of variables. Nine items have the negative *r*. This indicates that the larger values on the students' self assessment seemed to go with smaller values on the students' performance and vice versa. Put differently, those who said that they are good at understanding texts not related to their area of study obtained low marks in the questions related to that criterion and vice versa. Whether the *r* is positive, or zero or negative, the possibility that the correlation coefficient occurred by chance, in every item, is very high. This could be seen from the test significance at 2-tailed that all $p > .01$ or $p > .05$. For example, Question 15 has $p > .919$, which means more than 91% correlation coefficient happened by chance.

Table 5-17 below displays the *r* between the students' self assessment and their performance for the second criterion, understanding texts related to the area of study.

Table 5-17: The *r* for understanding texts related to the area of study

		READ6
VAR00007	**Pearson Correlation**	**.066**
	Sig. (2-tailed)	**.183**
	N	**413**
VAR00008	**Pearson Correlation**	**.131****
	Sig. (2-tailed)	**.008**
	N	**413**
VAR00009	**Pearson Correlation**	**.074**
	Sig. (2-tailed)	**.133**
	N	**413**
VAR00010	**Pearson Correlation**	**.054**
	Sig. (2-tailed)	**.274**
	N	**413**
VAR00011	**Pearson Correlation**	**.022**
	Sig. (2-tailed)	**.650**
	N	**413**
VAR00012	**Pearson Correlation**	**.045**
	Sig. (2-tailed)	**.363**
	N	**413**
VAR00026	**Pearson Correlation**	**.137****
	Sig. (2-tailed)	**.005**
	N	**413**
VAR00027	**Pearson Correlation**	**-.018**
	Sig. (2-tailed)	**.708**
	N	**413**
VAR00028	**Pearson Correlation**	**.057**
	Sig. (2-tailed)	**.252**
	N	**413**
VAR00029	**Pearson Correlation**	**-.028**
	Sig. (2-tailed)	**.575**
	N	**413**
VAR00030	**Pearson Correlation**	**.050**
	Sig. (2-tailed)	**.310**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

From Table 5-17, we note that the correlation coefficients between the candidates' self-assessment (READ6) and their performance (VAR0007-VAR00030) (see Appendix A.2.3: 471-74) are similar to the first criterion: very low. Except for two items (8 and 26), the correlation coefficients for all items are statistically not significant at either $p < .01$ or $p < .05$. All the interpretation relating to Table 5-16 above is relevant and sound when applied to the description of the findings in Table 5-17.

In Table 5-18 below, I display the correlation coefficient between the students' self-assessment (READ7-READ9) and their performance (TOTALRA-TOTALRC) for the third, fourth, and fifth criteria. Since the figures for these criteria are not as numerous as for the first two criteria above, I display the figures in one table.

Table 5-18: The *r* for answering questions in multiple-choice, true-false, and cloze formats

		READ7	READ8	READ9
TOTALRA	**Pearson Correlation**	**.259****	**.123**	**.238**
	Sig. (2-tailed)	**.000**	**.013**	**.000**
	N	**413**	**413**	**413**
TOTALRB	**Pearson Correlation**	**.254**	**.163****	**.217**
	Sig. (2-tailed)	**.000**	**.001**	**.000**
	N	**413**	**413**	**413**
TOTALRC	**Pearson Correlation**	**.250**	**.155**	**.268****
	Sig. (2-tailed)	**.000**	**.002**	**.000**
	N	**413**	**413**	**413**

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

From Table 5-18, we note that the correlation coefficients for the three criteria are still low: the *r* between READ7 (students' self-assessment) and TOTALRA (the total marks obtained by the students for Part A) is .26; the *r* between READ8 and

TOTALRB is .16; and the *r* between READ9 and TOTALRC is .27. However, these correlation coefficients are statistically significant at $p<.01$ as indicated by the (**) flag. In other words, we can be 99% sure that the correlation coefficients occurred for reasons other than chance.

5.5.1.1.2 The Grammar Test

A close scrutiny of the questionnaire reveals that from the total of eleven criteria, eight can be plotted with the test items to examine the correlation coefficient (see Table 5-19 below for the details of these criteria). With regard to the Grammar Test paper, several questions were chosen to be correlated with these criteria (see Appendix A.2.3: 475-79): five (32, 33, 38, 39, 42) for the first criterion; three (5, 21, 36) for the second; five (3, 15, 22, 23, 37) for the third; five (4, 9, 11, 24, 28) for the fourth; seven (7, 8, 10, 16, 26, 27, 41) for the fifth; four (12, 19, 29, 30) for the sixth; the total marks of Part A for the seventh; and the total marks of Part B for the eighth criterion. In Table 5-19 below, I display first the calculation of the means for the students' self-assessment and their performance in the Grammar Test for a comparative study. Since the scales for both variables differ from each other, the means are calculated in percentage.

Table 5-19: Means (in percentage) of self-assessment and performance for the Grammar Test

Criteria:	self-assessment (%)	performance (%)
1. Understanding *i`rāb*	64	39
2. Understanding *nakira* and *ma`rifa*	70	41
3. Understanding *mubtada'* and *khabar*	72	65

4. Understanding *kāna* and its sisters	68	39
5. Understanding *inna* and its sisters	67	45
6. Understanding *mufrad*, *muthannā*, and *jam`*	74	38
7. Answering the Grammar questions using the multiple choice format	66	46
8. Answering the grammar questions using the true-false format	64	70

We note from Table 5-19 that except for two criteria (3 and 8), the means for the students' self-assessment do not agree with their performance in the test: the means for their self-assessment are higher than their performance ranging between 22 and 30%. The students overestimated their actual ability especially in answering questions related to the grammar topics. With regard to the third and eighth criteria, we may say that the means for the students' self-assessment are, more or less, the same as their performance in the test; the difference is between 6 and 7% only.

To examine the relationship between the self-assessment and the performance of the same candidate, the variables were plotted using the SPSS programme. Tables 5-20 to 5-25 below display the correlation coefficient (*r*) of the test and the test significance based on the criteria described above.

Table 5-20: The *r* for *i`rāb*

		GRM1A
VAR00032	**Pearson Correlation**	**.139****
	Sig. (2-tailed)	**.005**
	N	**413**
VAR00033	**Pearson Correlation**	**.135****
	Sig. (2-tailed)	**.006**
	N	**413**
VAR00038	**Pearson Correlation**	**.049**

	Sig. (2-tailed)	**.321**
	N	**413**
VAR00039	**Pearson Correlation**	**.170****
	Sig. (2-tailed)	**.001**
	N	**413**
VAR00042	**Pearson Correlation**	**.082**
	Sig. (2-tailed)	**.098**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

From Table 5-20, we note that the correlation coefficient between the students' self-assessment (GRM1A) and their actual performance (VAR00032-42) on the *i`rāb* topic is very low, ranging between .05 and .17. This means that those students who thought they were good at *i`rāb* obtained low scores in the questions related to this topic or vice versa. With reference to the test significance, the *r* for three items is at $p < .01$. This figure indicates only 1% probability that the *r* for these three items occurred by chance alone. The *r* for the other two items is not significant at either $p < .01$ or $p < .05$.

Table 5-21: The *r* for *nakirah* and *ma`rifah*

		GRM1B
VAR00005	**Pearson Correlation**	**.057**
	Sig. (2-tailed)	**.249**
	N	**413**
VAR00021	**Pearson Correlation**	**.112***
	Sig. (2-tailed)	**.022**
	N	**413**
VAR00036	**Pearson Correlation**	**.084**
	Sig. (2-tailed)	**.089**
	N	**413**

* Correlation is significant at the 0.05 level (2-tailed).

From Table 5-21, we note that the *r* between the students' self-assessment of their ability (GRM1B) and their actual performance (VAR0005-VAR00036) on

nakirah and *ma`rifah* topics is very low. In terms of test significance, only the *r* between the GRM1B and VAR00021 is at $p < .05$.

Table 5-22: The *r* for *mubtada'* and *khabar*

		GRM1C
VAR00003	**Pearson Correlation**	**.105***
	Sig. (2-tailed)	**.032**
	N	**413**
VAR00015	**Pearson Correlation**	**.168****
	Sig. (2-tailed)	**.001**
	N	**413**
VAR00022	**Pearson Correlation**	**.116****
	Sig. (2-tailed)	**.018**
	N	**413**
VAR00023	**Pearson Correlation**	**.207****
	Sig. (2-tailed)	**.000**
	N	**413**
VAR00037	**Pearson Correlation**	**.031**
	Sig. (2-tailed)	**.528**
	N	**413**

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Table 5-22 shows that the *r* between the students' self-assessment of their ability (GRM1C) and their performance in the test, for *mubtada'* and *khabar*, is very low. Only the relationship between GRM1C and VAR00023 has the $r = .20$. However, four items have the *r* at either $p < .01$ or $p < .05$.

Table 5-23: The *r* for *kāna* and its sisters

		GRM1D
VAR00004	**Pearson Correlation**	**.198****
	Sig. (2-tailed)	**.000**
	N	**413**
VAR00009	**Pearson Correlation**	**.166****
	Sig. (2-tailed)	**.001**
	N	**413**
VAR00011	**Pearson Correlation**	**.242****
	Sig. (2-tailed)	**.000**
	N	**413**

VAR00024	**Pearson Correlation**	**.254****
	Sig. (2-tailed)	**.000**
	N	**413**
VAR00028	**Pearson Correlation**	**.118***
	Sig. (2-tailed)	**.016**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

In Table 5-23 above, we observe that the *r* between the students' self-assessment of their ability (GRM1D) and their performance in the test, for *kāna* and its sisters, is very low. However, the *r*'s for four variables are statistically significant at $p < .01$ and the *r* for another variable is at $p < .05$.

Table 5-24: The *r* for *inna* and its sisters

		GRM1E
VAR00007	**Pearson Correlation**	**.167****
	Sig. (2-tailed)	**.001**
	N	**413**
VAR00008	**Pearson Correlation**	**.182****
	Sig. (2-tailed)	**.000**
	N	**413**
VAR00010	**Pearson Correlation**	**.191****
	Sig. (2-tailed)	**.000**
	N	**413**
VAR00026	**Pearson Correlation**	**.070**
	Sig. (2-tailed)	**.158**
	N	**413**
VAR00041	**Pearson Correlation**	**.105***
	Sig. (2-tailed)	**.033**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 5-24 indicates that the *r* between the students' self-assessment of their ability (GRM1E) and their performance in the test for *inna* and its sisters is very low. Three variables have the *r* significant at $p < .01$ and one variable has the *r* at $p < .05$.

Table 5-25: The *r* for *mufrad, muthannā* and *jam`*

		GRM2A
VAR00012	**Pearson Correlation**	**.066**
	Sig. (2-tailed)	**.183**
	N	**413**
VAR00019	**Pearson Correlation**	**.092**
	Sig. (2-tailed)	**.061**
	N	**413**
VAR00029	**Pearson Correlation**	**.077**
	Sig. (2-tailed)	**.119**
	N	**413**
VAR00030	**Pearson Correlation**	**.081**
	Sig. (2-tailed)	**.101**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).

Table 5-25 shows that the *r* between the students' self-assessment of their ability (GRM2A) and their performance in the test for *mufrad, muthannā* and *jam`* is very low: less than 0.1 and not significant at either $p < .01$ or $p < .05$.

Table 5-26: The *r* for multiple-choice and true-false formats

		GRM3	GRM4
TOTALGA	**Pearson Correlation**	**.306****	**.320**
	Sig. (2-tailed)	**.000**	**.000**
	N	**413**	**413**
TOTALGB	**Pearson Correlation**	**.251**	**.269****
	Sig. (2-tailed)	**.000**	**.000**
	N	**413**	**413**

** Correlation is significant at the 0.01 level (2-tailed).

Table 5-26 displays the relationship for two criteria: the first refers to *r* between the students' self-assessment of their ability (GRM3) and their performance (TOTALGA) in answering the Grammar Test with the multiple-choice format; and the second refers to the *r* between their self-assessment (GRM4) and their performance (TOTALGB) in answering questions with the true-false format. The correlation coefficients for any of the criteria are relatively low: .31 and .27

respectively at $p < .01$.

5.5.1.1.3 The Dictation Test

From the total of eight criteria, I chose five only to examine the correlation coefficient (see Table 5-27 below for the details of these criteria). With regard to the students' performance, several questions were chosen: the total marks for the first criterion; four questions (1, 17, 21, 22) for the second; four (3, 5, 10, 14) for the third; two (4, 11) for the fourth; and lastly three questions (16, 24, 25) for the fifth criterion (see Appendix A.2.5: 505). Table 5-27 below summarises the means for the students' self-assessment and their performance in the Dictation Test for a comparative study.

Table 5-27: Means for the students' self-assessment and their performance for the Dictation Test

Criteria:	self-assessment (%)	performance (%)
1. Writing all items dictated	68	55
2. Determining whether the dictated word is one or more than one word	71	48
3. Writing words that have *alif lam Qamariyya*	76	62
4. Writing words attached to *alif lam Shamsiyya*	75	49
5. Determining the long and short vowels	70	31

From Table 5-27, we may suggest that the students overestimated their ability for the criteria listed above; their means of their self-assessment are higher than their performance in the test between 13 to 39%. To see the correlation coefficient for

those criteria, the scores of the students' self-assessment and their performance in the test was correlated. Table 5-28 to Table 5-30 below display the results of this correlation:

Table 5-28: The *r* for the first and second criteria: writing what was dictated and determining the words

		DICT1	DICT2
TOTALDIC	**Pearson Correlation**	**.288****	**.253**
	Sig. (2-tailed)	**.000**	**.000**
	N	**413**	**413**
VAR00001	**Pearson Correlation**	**.147**	**.066**
	Sig. (2-tailed)	**.003**	**.177**
	N	**413**	**413**
VAR00017	**Pearson Correlation**	**.115**	**.082**
	Sig. (2-tailed)	**.019**	**.094**
	N	**413**	**413**
VAR00021	**Pearson Correlation**	**.225**	**.175****
	Sig. (2-tailed)	**.000**	**.000**
	N	**413**	**413**
VAR00022	**Pearson Correlation**	**.079**	**-.051**
	Sig. (2-tailed)	**.109**	**.301**
	N	**413**	**413**

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

From Table 5-28 above, we note that the correlation coefficient between the students' self assessment (DICT1) and their performance (TOTALDIC) for the first criterion is low: .29, at $p < .01$. With regard to the second criterion, the correlation coefficient between the students' self-assessment (DICT2) and their performance in the test is very low too. The *r* between DICT2 and VAR00022 is -.05 The negative *r* indicates that those who thought they were good in determining whether the word that was dictated to them as one or more than one word obtained low marks in their actual performance and vice versa. However, the correlation coefficient between DICT2 and VAR00021 is at $p < .01$.

Table 5-29: The *r* for determining *alif lam Qamariyya*

		DICT3
VAR00003	**Pearson Correlation**	.125*
	Sig. (2-tailed)	.011
	N	413
VAR00005	**Pearson Correlation**	.065
	Sig. (2-tailed)	.188
	N	413
VAR00010	**Pearson Correlation**	.120*
	Sig. (2-tailed)	.014
	N	413
VAR00014	**Pearson Correlation**	.174**
	Sig. (2-tailed)	.000
	N	413

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

In Table 5-29, we note that the correlation coefficients between the students' self-assessment (DICT3) and their performance (VAR0003-VAR00014) for the third criterion are very low: less than .20. With reference to the test significance, two *r*'s are at $p < .05$ and one *r* is at $p < .01$.

Table 5-30: The *r* for two criteria: determining *alif lam Shamsiyya* (DICT4) and the long and short vowels (DICT8)

		DICT4	DICT8
VAR00004	**Pearson Correlation**	.179**	.185
	Sig. (2-tailed)	.000	.000
	N	413	413
VAR00011	**Pearson Correlation**	.196**	.119
	Sig. (2-tailed)	.000	.016
	N	413	413
VAR00016	**Pearson Correlation**	.078	.118*
	Sig. (2-tailed)	.113	.017
	N	413	413
VAR00024	**Pearson Correlation**	.128	.102*
	Sig. (2-tailed)	.009	.039
	N	413	413
VAR00025	**Pearson Correlation**	.126	.182**
	Sig. (2-tailed)	.011	.000
	N	413	413

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 5-30 includes the correlation coefficients for two criteria: the fourth and the fifth. The *r* between the students' self-assessment and their performance for either criterion, as shown in Table 5-30 above, is very low. The *r*'s for these variables are at either $p < .05$, (for two *r*'s), or $p < .01$, (for three *r*'s).

5.5.1.1.4 The Essay Test

All criteria in the questionnaire were chosen to examine the correlation coefficient. With regard to the students' performance, all except the total marks were chosen to be correlated with those criteria. The discussion starts with the calculation of the means for the students' self-assessment and their performance in the Essay Test for a comparative study.

Table 5-31: Means for the students' self-assessment and their performance for the Essay Test

Criteria:	self-assessment (%)	performance (%)
1. The content of the essay	62	47
2. The organisation of the idea	58	47
3. The use of grammar in writing an essay	48	54
4. The use of vocabulary in writing an essay	52	54
5. The use of punctuation eg. spelling etc.	66	57

We note from Table 5-31 that except for two criteria (3 and 4), the means for the students' self-assessment do not agree with their performance in the test: the means for their self-assessment are higher than the means for their performance, ranging between 9% and 15%. From the data in Table 5-31 above, we may say that

the students overestimated their actual ability, especially with regard to the criteria related to the content and organisation of the idea. With regard to the third and fourth criteria, we may say that the means for the students' self-assessment do, more or less, agree with their performance in the test. The differences are very small: 2% and 4% only.

To examine the relationship between self-assessment and performance, the variables were plotted using the SPSS programme. Table 5-32 below displays the correlation coefficient (r) of the test and the test significance based on the criteria described above. Since the variables for both self-assessment or performance are small in quantity, they are displayed in one table.

Table 5-32: The *r* for the criteria of the Essay Test

		ESSAY1	ESSAY2	ESSAY3	ESSAY4	ESSAY5
CTN.ORG	**Pearson Correlation**	**.082**	**.093**	**.147**	**-.005**	**.247**
	Sig. (2-tailed)	**.096**	**.060**	**.003**	**.925**	**.000**
	N	**413**	**413**	**413**	**413**	**413**
GRAMMAR	**Pearson Correlation**	**.067**	**.129**	**.136****	**.024**	**.194**
	Sig. (2-tailed)	**.177**	**.009**	**.006**	**.631**	**.000**
	N	**413**	**413**	**413**	**413**	**413**
VOCAB	**Pearson Correlation**	**.073**	**.101**	**.102**	**.022**	**.218**
	Sig. (2-tailed)	**.138**	**.040**	**.039**	**.657**	**.000**
	N	**413**	**413**	**413**	**413**	**413**
MECHANIC	**Pearson Correlation**	**.095**	**.071**	**.108**	**.036**	**.244****
	Sig. (2-tailed)	**.053**	**.151**	**.028**	**.466**	**.000**
	N	**413**	**413**	**413**	**413**	**413**

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

From Table 5-32 above, we note the correlation coefficients between the students' self-assessment (ESSAY1, ESSAY2, and ESSAY4) and their performance (CTN.ORG and VOCAB) for the first, second, and third criteria are very low: less than .1 and statistically not significant at either $p < .01$ or $p < .05$. We also note that

the correlation coefficients between the students' self-assessment (ESSAY3 and ESSAY5) and their performance (GRAMMAR and MECHAN) are also low but statistically significant at $p < .01$.

From the results of the students' self assessment above, we draw the following conclusions:

(A) The correlation coefficient between the students' self assessment and their actual performance in the criteria is generally very low. This indicates that the use of students' self-assessment to examine the concurrent validity of the placement test does not work in this case. The students, consciously or unconsciously, expressed opinions which did not always agree with their actual performance in the test. However, a very small number of them, as indicated in the result above, tended to have the same degree of assessment between their opinion and their performance.

(B) The statistics above show that the significance of the test ($p < .01$ or $p < .05$) does not necessarily indicate that the correlation coefficient should indicate a high correlation too. This means that the test may be statistically significant although the correlation coefficient is quite low. This probably explains what Rowntree (1991) means when he says that "...in the majority of cases, the question of 'satisfactoriness' [significance] is totally irrelevant" (p.170). However, the test significance proves that a very low or a negative correlation coefficient will cause the test to be not statistically significant.

(C) Since the correlation coefficients of most criteria in the test are low, we are obliged to accept the null hypothesis and therefore reject the alternative hypothesis for the variables with zero or negative correlation coefficients and not statistically significant at either $p < .01$ or $p < .05$. With reference to the variables with

positive correlation coefficients which were statistically significant at either $p < .01$ or $p < 05$, we are obliged to accept the null hypothesis too because the correlation coefficients were very low.

5.5.2 Parallel test result

The second measure of concurrent validity are the results of the examination which took place before the placement test was administered. I obtained the results for the examination by asking the students to write down their grades on the spaces provided in the questionnaire form. I have to stress here that with this method, the grades that were collected depended totally on what was written by the students. There was no counter check to verify the data.

If we refer to the original definition of concurrent validity, we find that the grades did not exactly parallel the test in terms of time. This is because the examination was administered either three or five months before the students matriculated at the university. Those who were admitted based on their results from the Higher Certificate of Education (HCE) took the Arabic paper for that examination in December. On the other hand, those who were admitted based on their results from the pre-AIS programme sat the Arabic paper at the end of the programme, usually at the beginning of March.

There are two main reasons why this type of test is valid, in examining concurrent validity of the placement test:

(a) the contents of these tests are, more or less, the same as the content of the placement test. The pre-AIS examination and the HCE certificate are the achievement tests: they assign grades, or certify mastery; while the current test is a

placement test: it determines entry performance. This means that those who scored good marks in the former test also tended to score good marks in the latter test and vice versa because either the achievement or the placement is a "...representative sample of course objectives" (Gronlund, 1982:19); and

(b) there was no learning activity between the first and the second test, i.e. between January and May or between April and May. This boils down to saying that the students depended on the same knowledge of Arabic to answer the questions for both test papers. The only difference which may have influenced the result is that the students took the achievement test immediately after they completed the course (the information was still fresh) while for the placement test they took it before they started the new course.

The number of papers and the skills tested for the HCE and the pre-AIS examination are summarised as follows:

(a) for the HCE, there were three papers: Paper 1 for oral skills; Paper 2 for grammar, writing and reading skills; Paper 3 for Arabic literature (*nuṣūṣ wa tārīkh al-adab*); and the total overall score for these papers (I classified this as Paper 4);

(b) for the pre-AIS examination, there were four papers: Paper 1 was for syntax and morphology skills; Paper 2 was for writing skills (essay); Paper 3 was for Arabic Rhetoric (*balāghah*); and Paper 4 was for Arabic literature (*nuṣūṣ wa tārīkh al-adab*).

The students were assigned grades from A to D for pass marks and F for fail marks. For the purpose of the data interpretation, I re-coded these grades into numbers as follows: 1 = D; 2 = C; 3 = B; and 4 = A. The frequency analysis indicated that no fail grade was disclosed. However, several students did not disclose their

grades: 51 for Paper One, 49 for Paper Two, and 42 for Paper Three. The most probable reasons are that they forgot their grades or they were unwilling to disclose their low grades. All of the blank spaces (no grades given) were assigned with number 1 (D) to avoid missing values in the data.

Before the correlation coefficient for these tests was examined, one null and one alternative hypothesis were set up. The null hypothesis for the second measure was as follows:

> There is no correlation coefficient between either the students' results for the HCE or the pre-AIS examination and their results for the placement test. Therefore the $r = 0.00$;

while the alternative hypothesis for the second measure was:

> There is correlation coefficient between either the students' results for the HCE or the pre-AIS examination and their results for the placement test.

Below is the description of the correlation coefficient of the results of both tests. The result of the Reading Test is discussed first followed by the Grammar Test, the Essay Test and lastly the Dictation Test.

5.5.2.1 The Reading Test

Table 5-33 displays the correlation coefficient between the Reading Test and both the results of the HCE certificate and the pre-AIS examination:

Table 5-33: The correlation between the Reading and the parallel tests

		TOTLREAD
P1RECODE	**Pearson Correlation**	**.601****
	Sig. (2-tailed)	**.000**
	N	**413**
P2RECODE	**Pearson Correlation**	**.448****
	Sig. (2-tailed)	**.000**
	N	**413**
P3RECODE	**Pearson Correlation**	**.339****

	Sig. (2-tailed)	**.000**
	N	**413**
P4RECODE	**Pearson Correlation**	**.583****
	Sig. (2-tailed)	**.000**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).

From Table 5-33, we note that the *r* between the Reading Test and Paper One, Two and Four are moderate: .60; .45; and .59 respectively at $p < .01$. The *r* between the Reading Test and Paper Three is quite low: .34. This is justifiable because Paper Three was related to Arabic literature and Arabic Rhetoric. It is normal for students, to the best of my knowledge, to obtain low marks for this paper.

5.5.2.2 The Grammar Test

Table 5-34 displays the correlation coefficient between the Grammar Test and the results of both the HSC certificate and the pre-AIS examination:

Table 5-34: The correlation between the Grammar and the parallel tests

		TOTALGRM
P1RECODE	**Pearson Correlation**	**.599****
	Sig. (2-tailed)	**.000**
	N	**413**
P2RECODE	**Pearson Correlation**	**.462****
	Sig. (2-tailed)	**.000**
	N	**413**
P3RECODE	**Pearson Correlation**	**.380****
	Sig. (2-tailed)	**.000**
	N	**413**
P4RECODE	**Pearson Correlation**	**.548****
	Sig. (2-tailed)	**.000**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).

From Table 5-34, we note that the *r* between the Grammar Test (TOTALGRM) and the grades for the three papers (P1, P2 and P4RECODE) are

moderate too: .60; .46; and .55 respectively, at $p < .01$.

5.5.2.3 The Essay Test

Table 5-35 below displays the correlation coefficient between the Essay Test and the results of both the HCE certificate and the pre-AIS examination:

Table 5-35: The correlation between the Essay and the parallel tests

		TTLESSAY
P1RECODE	**Pearson Correlation**	**.578****
	Sig. (2-tailed)	**.000**
	N	**413**
P2RECODE	**Pearson Correlation**	**.442****
	Sig. (2-tailed)	**.000**
	N	**413**
P3RECODE	**Pearson Correlation**	**.385****
	Sig. (2-tailed)	**.000**
	N	**413**
P4RECODE	**Pearson Correlation**	**.548****
	Sig. (2-tailed)	**.000**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).

The *r*'s between the total marks for the Essay Test (TTLESSAY) and the grades for four papers (P1RECODE-P4RECODE) do not differ very much from the first two tables: three correlation coefficients at moderate levels and one *r* relatively at $p < .01$.

5.5.2.4 The Dictation Test

Table 5-36 below displays the correlation coefficient between the Dictation Test and both the results of the HCE certificate and the pre-AIS examination:

Table 5-36: The correlation between the Dictation and the parallel tests

		TOTALDIC
P1RECODE	**Pearson Correlation**	**.606****
	Sig. (2-tailed)	**.000**
	N	**413**
P2RECODE	**Pearson Correlation**	**.406****
	Sig. (2-tailed)	**.000**
	N	**413**
P3RECODE	**Pearson Correlation**	**.385****
	Sig. (2-tailed)	**.000**
	N	**413**
P4RECODE	**Pearson Correlation**	**.554****
	Sig. (2-tailed)	**.000**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).

From Table 5-36, we note that the correlation coefficients between the total marks of the Dictation Test (TOTLDIC) and the grades for four papers (P1RECODE-P4RECODE) are as being interpreted earlier: three *r*'s are at a moderate level and one *r* is at a low level. All correlation coefficients are at $p < .01$.

From the data in Table 5-33 to Table 5-36 we draw the following conclusions:

(A) The positive correlation coefficient between the total marks of the placement tests and the grades for the HCE certificate and the pre-AIS examination indicates that there was a relationship between the students' performance in both tests. This relationship however was not very strong, as indicated by the moderate correlation coefficients. However; the correlation coefficient is statistically significant at $p < .01$. In other words, there is less than 1% probability that the correlation happened by chance only.

(B) The *r* between the total marks for the sub-tests (Reading, Grammar, Essay and Dictation) and Paper One (the oral skills for the HCE certificate and the grammatical skills for the pre-AIS examination) was the strongest one. On the other hand, the *r* between the total marks for the sub-tests and Paper Three

(*balāghah* and *nuṣūṣ*) was the lowest. It may be the case that the students were more familiar with the grammatical aspects of Arabic than literature and classical texts.

(C) The *r* between the results of these tests indicates that the interim between the two tests did not influence a close relationship. This finding may be used to argue that tests carrying the same construct can be used as an important criterion to examine concurrent validity even though the interval between the tests is three to five months, or even longer, as long as no learning activity takes place between the two tests.

(D) Since the correlation between both results is moderate and significant at $p < .01$, we are obliged to reject the null hypothesis.

5.6 Predictive validity

As discussed in Chapter Two (see 2.2.2.2.2), predictive validity involves the comparison of the test scores with some other measures for the same candidates taken some time after the test. Since the purpose of the placement test is to group students according to their ability leading to their achievement at the end of the course, it is thus important to investigate whether or not the placement test predicts the students' academic success.

The results that will be compared with the placement test (hereafter called the '*predictor*') are the results of the final examination in Arabic for Semester One, (hereafter called the '*sample*') which took place in October 1998. The predictive validity of the test would be the correlation coefficient (*r*) between the test results.

On the one hand, it is not likely that the correlation will be very high between the two tests. This is because the construction of the sample test was beyond our control in terms of validity, reliability, etc. Neither can we ascertain whether the sample test is in line with what Gronlund (1982), Harris (1988) and Brown (1996) suggest, i.e., a parallel test must be a test that is already a well-established measure of the construct involved. Another factor that may affect the correlation coefficient of both tests is that the language proficiency of the students may have improved as a result of learning activities which took place after the predictor test. The language instructors may also play an important role in improving the students' language proficiency. For example, they may have identified the weaknesses of their students from the predictor test and hence may have taken several steps to remedy those weaknesses. As a result, the students' achievement at the end of the course will be better than their achievement in the predictor test.

On the other hand, this is not always the case. In many instances, those who score low marks in the predictor test seem to be very unlikely to obtain very high marks in the next test and vice versa. This happens because some students are unable to cope with the remedial programme, which aims to increase their ability in particular aspects of language. Another reason is related to the nature of the learning itself, which does not focus on the remedial programme; instead the focus is on completing the syllabus.

5.6.1 The content and descriptive statistics of the sample test

Before the correlation coefficient is conducted, the section below attempts to describe briefly the content and the descriptive statistics of the sample test.

(i) The content:

The content of the test paper, which was prepared by teachers at the Faculty of Languages and Linguistics, can be summarised as follows (for the details of the examination paper, see Appendix A.2.6: 507-523):

(a) Cover page. The Malay and Arabic rubrics instruct the students on how to answer the questions; the time allocated is two hours and a half. There are no examples of how to answer the questions, perhaps due to the variety of question-types contained in the test.

(b) Test contents. The test consists of three parts: Part One relates to syntax, morphology, and Arabic Rhetoric (39 multiple-choice questions) (see Appendix A.2.6: 507-19); Part Two consists of translation from and to Arabic (10 multiple-choice questions) (see Appendix A.2.6: 519-23); and Part Three involves writing an essay on given topics (see Appendix A.2.6: 523). The total mark is divided into two parts only: the total mark for Part One and Part Two is sixty and the total mark for Part Three is fifteen. This gives a total of seventy five ($k = 75$) altogether.

(ii) Descriptive statistics

Table 5-37 below displays the descriptive statistics for the sample one test including central tendency and dispersion.

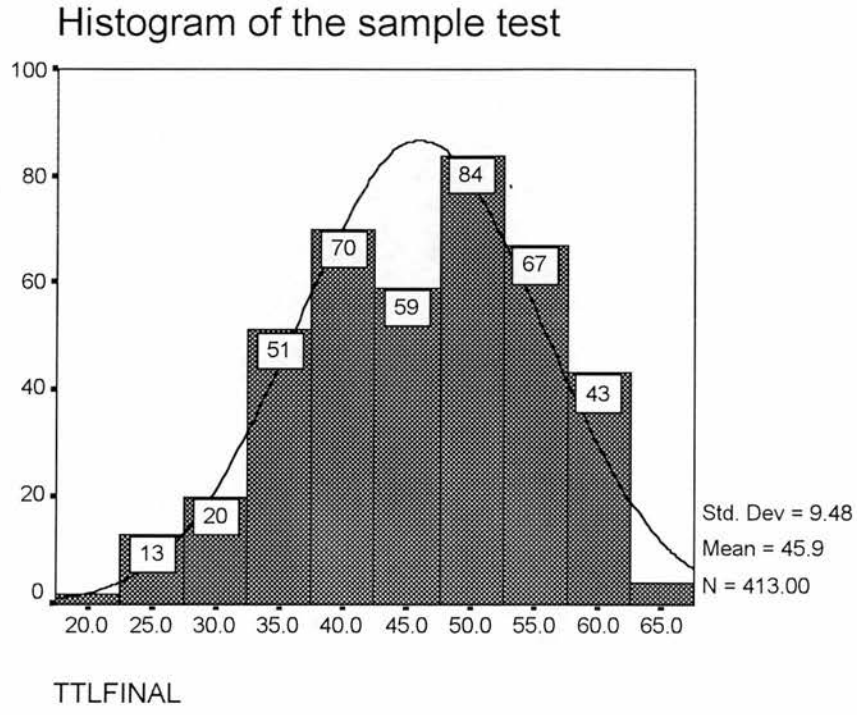
Table 5-37:Descriptive statistics for the sample paper

N	**Valid**	**413**
	Missing	0
Mean		45.9346
Median		47.0000
Mode		52.00
Std. Deviation		9.4821

Variance		89.9108
Range		46.00
Minimum		20.00
Maximum		66.00
Sum		18971.00

From Table 5-37, we note that the mean, 45.93, seems at first glance to suggest that the students found the test slightly easy as the mean is above the 50% mark (precisely 61.24%), and therefore there are probably more candidates situated towards the top end of the distribution than the bottom end. The median, 47.0, and the mode, 52.0 also indicate that the majority of the students obtained high marks. With reference to the dispersion, the standard deviation for the test, 9.5, indicates that we can fit 2 SDs (55.4 for 1 SD and 64.9 for 2 SD) only on the + (positive) side of the mean which would account for 95% of the population and nearly 3 SDs (36.4 for 1 SD, 26.9 for 2 SD, and 17.4 for 3 SD)on the - (negative) side of the mean which would account for 99.7% of the population. This distribution is negatively skewed and therefore confirms the suggestion above that most of the candidates obtained high scores on the sample test. To give a clearer picture of the distribution of samples for the test, I display the histogram in Figure 5-11 below:

Figure 5-11: Histogram of the sample test



The histogram in Figure 5-11 shows that the distribution is negatively skewed. We can see that the ends of the normal curve line disappear off the histogram before 20 at the negative side and not exactly at 65 at the positive side (3 SD is 74.4). This reflects the figures that fit the distribution I calculated above concerning 2 SDs on the positive side and nearly 3 SDs on the negative side, i.e., 64.9 and 17.4.

5.6.2 Correlation analysis between the predictor and the sample

This section attempts to analyse the relationship between the total scores of the placement test (the predictor) and the total scores of the final examination for semester one (the sample). Before the correlation analysis is conducted, one null (H_0) and one alternative (H_1) hypothesis were set-up as follows:

> For (H_0): There is no correlation coefficient between the total score of the *predictor* and the total score of the *sample*. Therefore the r is = 0.00; and
> For (H_1): There is correlation coefficient between the total score of the *predictor* and the total score of the *sample*. Therefore the r is > 0.00.

Table 5-38 below displays the correlation between the total marks for the predictor and the total marks for the sample test:

Table 5-38: The *r* between the total scores of the predictor and the sample

		TTLFINAL
TTLREAD	**Pearson Correlation**	**.595****
	Sig. (2-tailed)	**.000**
	N	**413**
TTLGRAM	**Pearson Correlation**	**.639****
	Sig. (2-tailed)	**.000**
	N	**413**
TTLESSAY	**Pearson Correlation**	**.601****

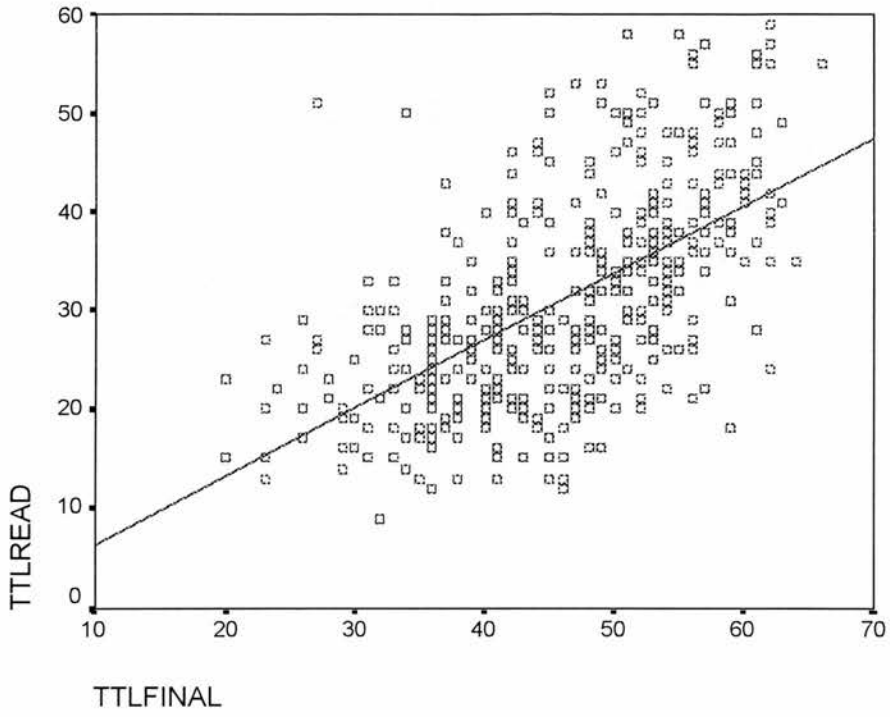
	Sig. (2-tailed)	**.000**
	N	**413**
TTLDIC	**Pearson Correlation**	**.613****
	Sig. (2-tailed)	**.000**
	N	**413**

** Correlation is significant at the 0.01 level (2-tailed).

With reference to the first row in Table 5-38 above, we note that the *r* between the total score of the sample (TTLFINAL) and the score of the Reading Test (TTLREAD) is .60 (to two decimal points) at $p < .01$. With reference to the fourth row, the *r* between TTLFINAL and the total score of the Grammar Test(TTLGRAM) is .64 at $p < 01$. In the seventh row, the *r* between TTLFINAL and the total scores of the Essay Test (TTLESSAY) is = .60 at $p < 01$. From these findings, we may say that predictive validity for the placement test can be described as moderate as indicated by the correlation coefficients of both tests.

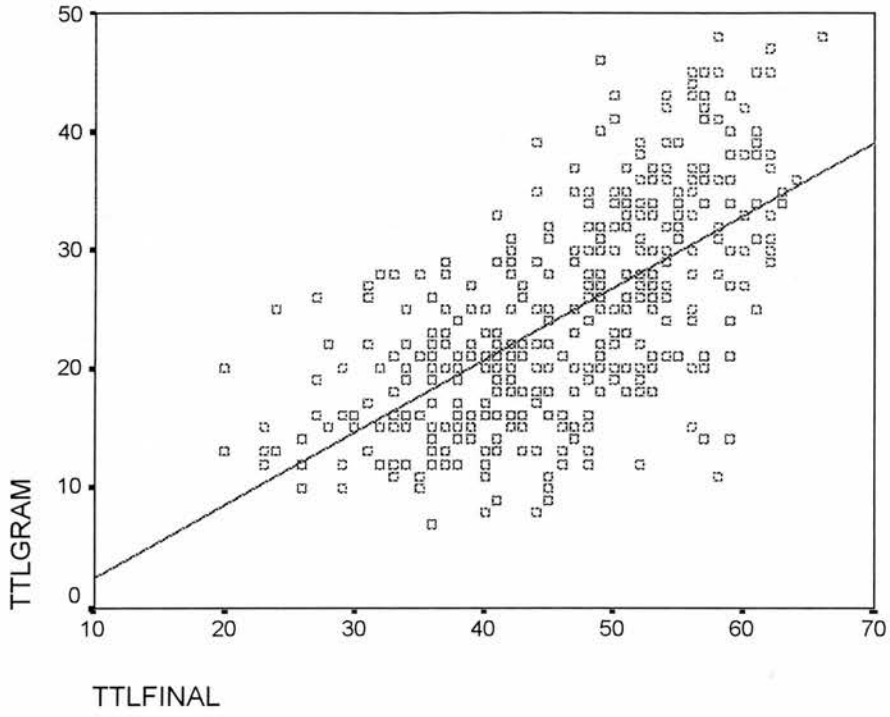
To get a clearer picture of the relationship between the predictor and the sample tests, the scores were plotted using the SPSS programme. The description of the scatterplot for the relationship between the sample and the Reading Tests is displayed first followed by the Grammar, the Essay and the Dictation. Figures 5-12 to 5-15 below display the outcomes of the scatterplot:

Figure 5-12: Scatterplot for the Reading and Sample Tests



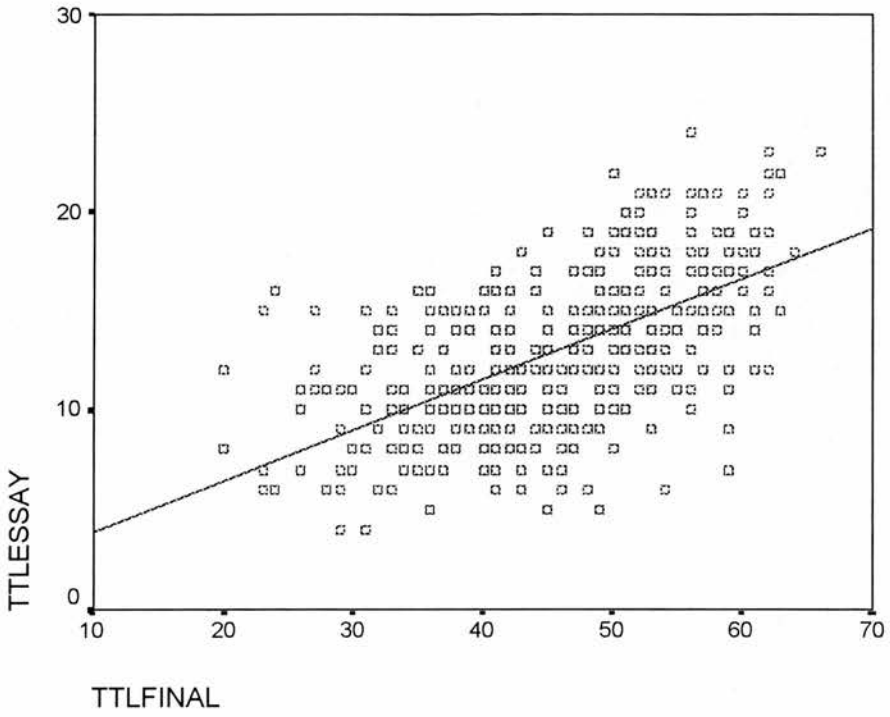
From the shape and slope of the squares in Figure 5-12, we can see clearly that the type of scatterplot for this relationship is a positive one, i.e. larger values on the sample test go with larger values on the Reading Test, and vice versa. The relationship however is not very strong because the plotted points on the diagram do not lie close to the straight line. In addition, quite a large number of points are scattered, indicating that some differences are likely between the scores: a number of students who obtained high scores in the sample test scored low marks in the Reading Test, which contributed to the moderate correlation coefficient of the two tests.

Figure 5-13: Scatterplot for the Grammar and Sample Tests



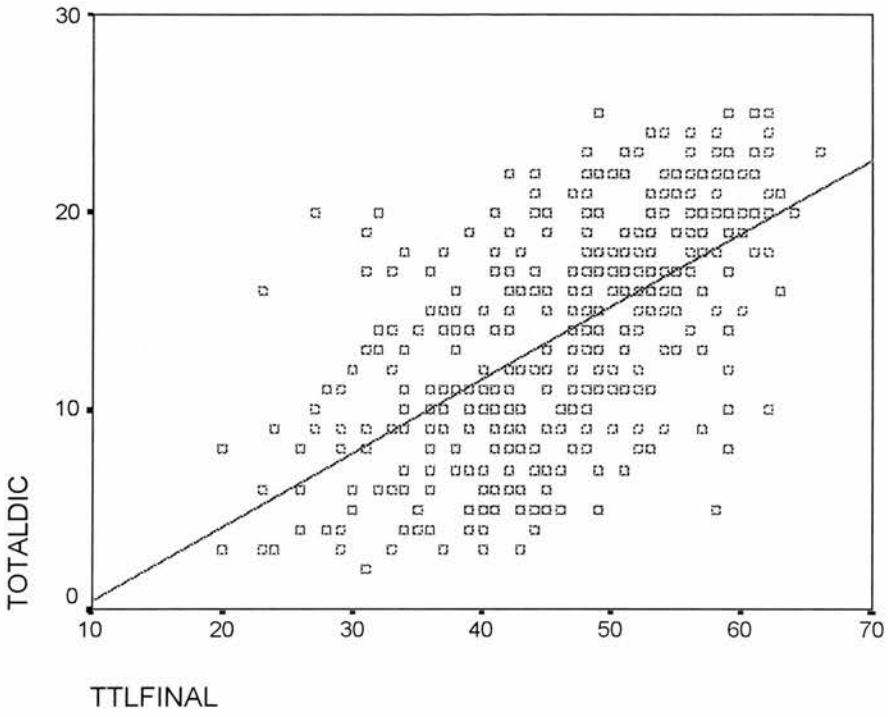
The slope and shape of squares in Figure 5-13 indicate that the type of scatterplot for this relationship is a positive one too. The plotted points on the scatter diagram that lie on the straight line are more concentrated in this diagram than the plotted points in Figure 5-12 above. However, the relationship is still considered moderate. A small number of points are located away from the straight line, especially between 20 and 40 for the Y axis and between 30 and 60 for the X axis, indicating the likelihood of some differences between the scores. It is also noted from Figure 5-13 above that a small number of students who scored high marks for the sample test (between 52 to 60) obtained very low marks for the Grammar Test (between 10 and 12 marks only).

Figure 5-14: Scatterplot for the Essay and Sample Tests



The slope and shape of squares in Figure 5-14 indicate that the type of scatterplot for this relationship is a positive one. It is also observed from Figure 5-14 above that, as in Figure 5-13, some students who scored high marks in the sample test (between 50 and 60), obtained very low marks in the Essay Test, ranging between 4 and 8 marks. This may be one of the reasons why the correlation coefficient of both tests arrives at the moderate level only.

Figure 5-15: Scatterplot for the Dictation and Sample Tests



The slope and shape of squares in Figure 5-15 indicate that the type of scatterplot for this relationship can be described as a positive one too. The relationship, however, is not very strong because the plotted points on the diagram, as described in the figures above, do not lie close to the straight line. In addition, quite a large number of points are scattered, indicating that some differences are likely between the scores: a number of students who obtained high scores in the sample test scored low marks in the Dictation Test. For example, a number of students who obtained scores between 40 and 60 for the sample test scored very low marks for the Dictation: between 3 and 10 marks only, which may contribute to the moderate correlation coefficient of the two tests.

From the correlation analysis between the scores of the placement test (the

predictor) and the scores of the final examination (the sample) together with the distributions of the relationship between these two scores which were displayed in the form of scatterplot, we draw the following conclusions:

(i) The correlation coefficient between the scores of the two tests was at the moderate level at $p < .01$. We may suggest here that the predictor's scores for every student would anticipate the students' achievement at the end of the course, if both tests are to be conducted in the future on different candidates but at a similar level of language proficiency.

(ii) External factors may contribute to the moderate level of the correlation coefficient between both tests. Among these factors is, as stated in 5.6 above, the language proficiency of the students, which may have improved as a result of a learning activity which took place after the predictor test. It can be seen from the descriptive statistics in Figure 5-11 above that many students scored higher marks than lower marks. This boils down to saying that the correlation analysis between the placement and the achievement tests, in many instances, may result in low and moderate levels of correlation coefficient.

(iii) Since the correlation coefficients between the predictor and the sample test are moderate (not 0.00 correlation) and are significant at $p < .01$, I am obliged to reject the null hypothesis.

5.7 Summary of Chapter Five

This chapter has examined the external analysis of the tests, preceded by the administration of the test at AIS, the investigation of the content validity, the

reliability and the correlation of the test. Evidence was presented for the existence of face and content validity (from the teachers' perspective) in the placement test. In the light of the reliability, the analysis indicated that the placement test is very reliable: the reliability coefficient (r_{xx}) ranged between .87 and .90 for the Reading, Grammar, and Dictation Tests and the rank order correlation (ROC) ranged between .74 and .89 for the Essay Test. As for the correlation analysis, the correlation coefficient (r) showed that the sub-tests are relatively correlated to each other: the r ranged between .69 and .75. Concurrent validity was obtained using two different types of measures: the students' self-assessment and the parallel test result. The analysis between both measures and the placement tests indicated that with the first measure, the correlation coefficient was very low while with the second measure, there is an association between both tests. So far as the predictive validity is concerned, most significant coefficients were above .59, reaching a maximum of .64. We may suggest at the end of this chapter that the placement test is a valid measure of the Arabic proficiency of the new students at the AIS.

6. CHAPTER SIX: SETTING PASS MARKS, CONCLUSION AND RECOMMENDATION

6.1 Introduction

The task of the test construction is not complete until pass marks have been set by the test constructor. In this chapter, I will start the discussion with the setting of pass marks. This will be followed by my conclusions and the recommendations stemming from this research.

6.2 Setting pass marks

There are various ways of setting pass marks. According to Alderson *et al.* (1996), some of the methods of setting pass marks, as practiced by individual test constructors or test boards, are: (i) *a fixed percentage* such as a 50% or 60% as a pass mark etc.; (ii) *a grade on the curve*, which refers to the normal distribution "...and assumes that normal distributions occur and are appropriate distributions for language proficiency and learning" (op. cit: 156); (iii) identify "...*'masters'* - people who are known to possess the ability being measured such as native speakers who can competently use the language on which the candidates are being tested - and see how well they perform on the test" (op. cit: 157); and (iv) "a *'standard setting'* where trained professionals with relevant expertise inspect the content of the test and then decide what the likely performance of *barely-adequate* candidates on this test would be" (p.158).

With regard to this study, I have chosen to use the first two procedures, i.e. a fixed percentage and a grade on the curve, when setting pass marks for the four sub-tests, i.e. Reading, Grammar, Essayand Dictation. The reason for selecting the former procedure is that this is the normal practice at the AIS in setting pass marks for its students for any examination or test conducted on the students. The reason for selecting the latter is that this procedure is in line with my current research which concentrates on the use of statistical data, including the curve, normal or positively or negatively skewed distributions, etc. The discussion below starts with the fixed percentage procedure and continues by looking at the grade on the curve procedure.

6.2.1 The fixed percentage procedure

With this procedure, the examiner has to use the fixed percentage which has been assigned by the examination board of the institution where the test takes place. The scales, which are used by the Examination board at the University of Malaya, are as follows:

Marks (in %)	Grades	Criteria
39 and below	Fail	very weak
40-44	D	weak
45-49	D+	weak
50-54	C	average
55-59	C+	average
60-64	B	very good
65-69	B+	very good
70 and above	A	Excellent

To make the total mark fit the above scale, I converted the total scores of every sub-test into a percentage: the total mark for every candidate divided by the total mark of the sub-test multiplied by one hundred. Below are the summaries of the grades, percentages, and total number of candidates for every sub-test of the

placement test:

1. The Reading Test

Grades	Percentage	Total candidate (of N=413)
Fail	52.1%	215
D-D+	20.3%	84
C-C+	13.3%	55
B-B+	10.7%	46
A	3.1%	13

2. The Grammar Test

Grades	Percentage	Total candidate (of N=413)
Fail	32.2%	132
D-D+	22.3%	91
C-C+	16%	66
B-B+	14%	58
A	15.5%	66

3. The Essay Test

Grades	Percentage	Total candidate (of N=413)
Fail	20.8%	86
D-D+	27.9%	115
C-C+	14.7%	60
B-B+	21.1%	86
A	15.5%	66

4. The Dictation Test

Grades	Percentage	Total candidate (of N=413)
Fail	28.3%	115
D-D+	15.5%	66
C-C+	9%	36
B-B+	16.4%	69
A	30.8%	127

6.2.2 The grade on the curve procedure

With this method, the grade assigned to the candidates depends on the distribution of the marks in that particular test. For the purpose of setting pass marks for the candidates at the AIS, I summarise below the distribution of marks together with the means, and the frequency of the standard deviation (SD) of the distribution for every sub-test (for the details of the distribution, see Chapter 5: 5.2.3):

1. For the Reading Test (total mark = 75):

the mean:	**the distribution of marks**
31.01 (41%)	(+) side : nearly 3 SDs (41.89, 52.78, 63.67)
	(-) side: 2 SDs (20.11, 9.22)

2. For the Grammar Test (total mark = 50):

the mean:	**the distribution of marks**
24.37 (49%)	(+) side : nearly 3 SDs (33.4, 42.4, 51.4)
	(-) side: 2 SDs (15.4, 6.4)

3. For the Essay Test (total mark = 25):

the mean:	**the distribution of marks**
13.06 (54%)	(+) side : nearly 3 SDs (17.04, 21.07, 25.1)
	(-) side: 2 SDs (9.07, 5.06)

4. For the Dictation Test (total mark = 25):

the mean:	**the distribution of marks**
13.72 (55%)	(+) side : 2 SDs (19.42, 25.12)
	(-) side: 2 SDs (8.02, 2.32)

To give a grade to candidates based on the standard deviations, Alderson *et al.* (op. cit: 156) suggest that:

> "Those who are more than, say, two standard deviations above the mean may be considered to be 'excellent', or 'exceptional', and receive the highest grade; those

scoring between one and two standard deviations above the mean are considered to be 'good' and classified accordingly; and so on down to 'exceptionally weak' for those whose score falls more than three standard deviations below the mean".

Alderson, *et al.* add that a given score expressed in the standard deviations is not necessarily in harmony with the standard pattern, simply because there are sometimes more or fewer than three standard deviations above or below the mean. Since some distributions of marks in the sub-tests described earlier are less than three standard deviations, either above (+) or below (-) the mean, I will assign a grade to the candidates according to their scores expressed in terms of standard deviations. The grades for the Reading Test are described first followed by the Grammar Test etc. (The displayed grades include the percentage, the total number of candidates who obtained that grade and the standard deviations (SD) in which the grades are assigned)

1. The Reading Test

Grades	**Percentage**	**total candidates**	**SD (+ or -)**
Fail	18.4	76	2 SD (-)
D	38.7	160	1 SD (-)
C	24.3	100	1 SD (+)
B	15.5	64	2 SD (+)
A	3.1	13	above 2 SD (+)

2. The Grammar Test

Grades	**Percentage**	**total candidates**	**SD (+ or -)**
Fail	18.6	77	2 SD (-)
D	35.9	148	1 SD (-)
C	27.3	113	1 SD (+)
B	14.3	59	2 SD (+)
A	3.9	16	above 2 SD (+)

3. The Essay Test

Grades	Percentage	total candidates	SD (+ or -)
Fail	20.8	86	2 SD (-) and below
D	34.4	142	1 SD (-)
C	29.3	120	1 SD (+)
B	14	59	2 SD (+)
A	1.5	6	above 2 SD (+)

4. The Dictation Test

Grades	Percentage	total candidates	SD (+ or -)
Fail	16.7	69	2 SD (-) and below
D	27.1	112	1 SD (-)
C	36.1	149	1 SD (+)
B	17.9	74	2 SD (+)
A	2.2	9	above 2 SD (+)

It is important to note here that the procedures used will depend to a large extent upon what the purpose of the test is (Alderson, *et al.* op. cit.). That said, we may suggest here that the second procedure, 'a grade on the curve', seems to be workable for the purpose of grouping students for placement according to their ability. This is so for the following reasons:

(A) The number of candidates obtaining the grades for different sub-tests in the second procedure is consistent compared with the first procedure. For example, the number of candidates obtaining the lowest and the highest grade (Fail and A) for these sub-tests is about the same.

(B) The second procedure assigns grades to the candidates based on what the candidates obtained in the test, while, with the first procedure, the candidates are graded according to grades that have been fixed. Therefore, we note that many students (more than 50%) fail the Reading, Grammar, and Dictation Tests when we assign grades using the first procedure.

(C) With the second procedure, we can decide whether or not to exempt high achievers from the course or to have a remedial course for some candidates. This is because the number of candidates situated at the three standard deviations above or below the mean is normally very small compared to the number of the candidates at other standard deviations. This, however, cannot be done with the grading marks in the first procedure because too many candidates scored A grade except for the Reading Test.

(D) The distribution of candidates by marks in the second procedure is equally divided as a result of refining and modifying the test items. This is not the case when we use the first procedure. The distribution of marks is not even especially with grades A and F.

Taking these factors into account, we may suggest that for any placement test, marks may need to be set using the grade on the curve procedure, in making a decision on the candidates. This suggestion, however, does not ignore the important roles that the fixed percentage procedure plays in setting pass marks in other types of test such as achievement tests etc.

6.3 Conclusions

The primary aim of this research is to construct and to validate an Arabic placement test for the use of new intake students in the Academy of Islamic studies at the University of Malaya. With reference to the first part of this research, the construction of the test, several steps were executed before this task could be undertaken, to ensure the task meets the requirements in the area of language testing.

The first step was an investigation of trends in language testing, as a literature review to the research. This review was conducted to guide the researcher in choosing between the trends in the development of measuring instruments. The literature review threw up two important factors: trends in language testing, and the influence of the approaches of teaching on language testing. With regard to the former, three major trends in language testing were noted: pre-scientific, psychometric structuralist, and socio-linguistic. As for the latter, it was observed that the design of language testing tends to follow a language teaching approach. As a result, the construction of the present test adopted some of these trends without ignoring the importance of the influence of language teaching approaches practiced at the Academy of Islamic studies (AIS).

The second step involved the analysis of current tests together with the syllabus used at the AIS to obtain information for the construction of a new test for this research. It was found that some tests items have a very low content and face validity because they are not related to the syllabus. It was also found that the test constructors at the AIS did not prepare the test specification nor the investigation of validity and reliability of the test . This provides the researcher with an opportunity to construct a test that is based on the syllabus and the test specifications outlined in Chapter Two and Chapter Three of this thesis respectively. In addition, an investigation of the placement and proficiency tests, as indicated in Chapter Three, from various countries such as Saudi Arabia, America, Jordan, and Malaysia, has helped the researcher to design the placement test for this research.

This research has also established that the issue of constructing a good test does not rely solely on the assurance of a close relationship between the test and the

syllabus. The pilot study proved that many items, which are related to the syllabus, have low indices for either item facility (IF) or item discrimination (ID) or distractor efficiency or for all of these statistical measures. Therefore, special consideration was given to the statistical measures as well as the syllabus before any decision was taken as to whether or not to discard or to retain the test items.

The selection of the respondents in this research could be considered successful. The involvement of native speakers of Arabic and respondents with a higher level of Arabic, as with the first group of respondents for example, helped the researcher to determine the difficulty level of items. For example, the analysis of text one in Part Three for the Reading Test revealed that the majority of the respondents did not answer correctly the questions in this part. Since the academic level of the respondents was very much higher than the target samples in the real test, it was anticipated that the candidates in the real test would not be able to answer the questions in this part. As a result, the text was removed before the second part of the pilot study was conducted. The general conclusion, which the researcher used to scrutinise items with a high level of item facility (difficulty), seemed to work well for the purpose of obtaining higher content and face validity (see Chapter Four, 4.7.1.3.2 for the details of the general conclusion). The same applies to the respondents with a lower level of academic background than the candidates in the real test, as with samples from secondary schools in Malaysia. The researcher could easily identify items with high index of item facility (IF) or items with low index of item discrimination (ID) as an easy item and could therefore consider modifying them or discarding them from the final version of the test.

With regard to the second part of the study, the validation, the results of internal and external analyses of the test indicate that this study has successfully achieved its goal. All except one null hypothesis set up in Chapter Five have been rejected. This means that the alternative hypotheses, which are the main target of this research, are statistically and empirically acceptable. Below is the summary of the results of the internal and external analyses, bringing this research to a successful conclusion:

(A) As for internal validity, the construction of the test, based on the syllabus at the AIS, shows that all items for sub-tests are related or at least loosely related to the syllabus. The descriptive and item analyses, which include central tendency and dispersion for the former, and item facility, item discrimination and distractor efficiency for the latter, detected items with low indices which were removed from the final version of the test. A further investigation of face and content validity, made by the teachers at the AIS, also provided a very strong evidence of the high content and face validity of the test: the majority of teachers gave a verdict of between 1 and 2 (very good and good) on the four-point scale.

(B) As for the reliability analysis, the Reliability Coefficient (r_{xx}), of three sub-tests was very high: between .87 and .90. As for the Essay Test, the rank order correlation (ROC) of this sub-test was between .75 and .89. This high reliability coefficient of the test items shows the consistency of the test. To prove this consistency, the researcher conducted another test on 555 new students at the AIS for the session 1999/2000 last June. Two sub-tests were used for this research: the Reading and the Grammar (see Appendix A.2.4: 483-494 for the test items and see Appendix B.3: 544-545 for the descriptive statistics of the total scores of these two

papers). The Dictation and the Essay Tests were not conducted due to time constraints. The findings show that the test is consistent, i.e. reliable. Table 6-1 below summarises the results of both tests for comparative study:

Table 6-1: The reliability of the test items

Years	1998/99(N=413)	1999/2000(555)
(i) The Reading Test:		
Part One		
the mean	6.4	6.4
the SD	2.14	2.17
Part Two:		
the mean	12.2	11.4
the SD	2.98	2.82
(ii) The Grammar Test:		
Part One:		
the mean	20.8	19.7
the SD	8.27	8.37

(Note: Part Three of the Reading and Part Two of the Grammar Tests are not shown here because some changes were made to the items in the second trial)

(C) With reference to the correlation analysis, the correlation coefficient (r) between the four sub-tests was no less than .70 and significant at $p<.01$. This indicates how closely the two sets of scores correspond. It also indicates that the correlation occurred for reasons other than chance; the possibility of the relationship occurring by chance is less than one percent. This high correlation is very important, if the high validity of the test is to be ensured. If the correlation coefficient was low for example, this means that the two sets of tests examined different constructs.

(D) With regard to the analysis of external validity, the analysis of concurrent validity gave a convincing result. Even though the majority of the means for the criteria in the students' self-assessment were low, the correlation coefficient between parallel test results and the placement test was moderate. Some factors may influence the low means in students' self-assessment. Among these factors was that the students seemed not to be familiar with the technique of assessing themselves. Therefore,

we notice earlier in the analysis that they overestimated or underestimated their actual ability in the test, etc. As far as predictive validity goes, the result of the analysis showed that even though the correlation coefficient between the placement test and the final examination was moderate, the scatterplot of all sets of scores was always a positive one. Put differently, there was a 'go-togetherness' between the two sets of scores of the two tests.

6.4 Recommendations for future research

Before putting forward suggestions for future research, it is worth mentioning here the problems encountered during this research. There were two main problems:

(A) The researcher exercised no control over the teaching of Arabic at the AIS for the period between June and September (First semester). After the result of the placement test was delivered to the Arabic instructors at the AIS, there was no follow-up to ensure that the students were placed in a group according to their proficiency.

(B) The researcher had no control over the construction of the final examination for Arabic for the first semester.

It is suggested therefore for future research that these two aspects should be addressed. With regard to the first problem, it is important to monitor and in some cases to interfere with the approaches used in the teaching of Arabic in classes, the courses designed for students, the skills focused on in every group, the teaching materials used for the course, etc., so that the researcher is aware of the students' overall improvement. All of these are to ensure that the teaching of Arabic is

appropriate to the students' standard in Arabic. If, for example, the students need a remedial course, the course designed should suit their needs. The same applies to those who obtained excellent results in the placement test: they may need to be exempted from the course and offered an advanced course. Only then can the effort of conducting a placement test be considered worthwhile.

With reference to the second problem, it is very important to have some control over the construction of the final examination in Arabic, i.e. the achievement test. This is to ensure that the format, the content, and the skills tested, are in line with the format of the placement test. All this is for the purpose of predictive validity. If the final examination paper in Arabic for the first semester (1998/99) consisted of the four sub-tests, i.e. the Reading, Grammar, Writing, and Dictation, then the findings of predictive validity would be more meaningful. We could see clearly whether the correlation coefficient (r) between the scores of the sub-tests is high or low and whether these scores are significant or not. Moreover, the analysis can also be conducted on every sub-test, e.g. the Reading Test for the placement test against the Reading paper for the final examination, etc. Last but not least, in terms of administrative work, the involvement of a researcher in teaching and preparing questions will give him or her easy access to the data. This researcher's experience of being denied access to the results of the Arabic papers for the second semester examination by the administration at the AIS is a good illustration of the practical problems caused by a lack of access to all necessary data.

Bibliography

Aiken, L.R. (1964). Item context and position effects on multiple-choice test. *Journal of Psychology*, 58, 369-373.

Aitken, K.G. (1977). Using cloze procedure as an overall language proficiency test. *TESOL Quarterly*, 11. 1: 59-67.

Alderson, J.C. (1978). *A study of the cloze procedure with native and non-native speakers of English*. Unpublished Ph.D thesis. University of Edinburgh.

_____ (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13. 2: 219-227.

_____ (1988). New Prosedures for validating proficiency tests of ESP? Theory and practice. *Language Testing,* 5, 220-32.

_____ (1990). Testing reading comprehension skills (Part One). *Reading in a Foreign Language*, 6, 425-38.

_____ (1990). Testing reading comprehension skills (Part Two) : Getting students to talk about taking a reading test (A pilot study). *Reading in a Foreign Language*, 7, 465-502.

_____, Clapham, C. and Wall, D. (1996). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

_____, and Lukmani, Y. (1989). Cognition and Levels of Comprehension as Embodied in Test Question. *Reading in a Foreign Language*. 5(2): 253-270.

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing*, 9, 101-122.

Allen, V. F. (1968). Towards a thumbail test of English competence. *TESOL Quarterly*, 2. 241-246.

Allen, J.P.B., & Davies, A. (Eds.). (1977). *Testing and experimental methods*. The Edinburgh Course in Applied Linguistics (Vol. 4). London: Oxford University Press.

Allen, M.J. ,& Yen, W.M. (1979). *Introduction to measurement theory*. Monterey, Calif.: Brooks/Cole.

American national standards institute. (1977). *American national standard for bibliographic references.* New York.

American Psychological Association. (1996). *Publication manual of the American Psychological Association.* Fourth Edition. Washington DC.

Anastasi, A. (1988). *Psychological Testing*. London: McMillan.

Anderson, D. F. (1953). Test of achievement in the English language. *English Language*

Teaching, 7. 2: 37-69.

Anderson, J. (1970). A technique for measuring reading comprehension and readability. *English Language Teaching*, 25. 1: 178-182.

_____ (1971). Selecting a suitable 'reader': Procedures for teachers to assess language difficulty. *RELC Journal*, 2. 2: 35-41.

Anderson, N.J. , Bachman, L. , Perkins, K. , & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources: *Language Testing*, 8, 41-46.

Arena, L.A. (1990). *Language Proficiency: Defining, teaching, and testing*. Plenum Press, New York.

Bachman, L.F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, 16, 61-70.

_____ (1990). *Fundamental consideration in language testing*. Oxford: Oxford University Press.

_____ (1991). What does language testing have to offer? *TESOL Quarterly*, 25, pp. 671-704.

_____, Lynch, B. K. , & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12, 238-257.

_____, & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, 6, 14-29.

_____, & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Baker, R. L. (1987). *An investigation of the Rasch Model in its application to foreign language proficiency testing*. Unpublished PhD thesis. University of Edinburgh.

Bennet, W. A. (1968). *Aspects of Language and Language Teaching*. London: Cambridge University Press.

Berk, R. A. (Ed.). (1984). *A Guide to Criterion-Referenced Test Construction*. Baltimore, Md. : The Johns Hopkins University Press.

Blais, J. -G., & Laurier, M.D. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing*, 12, 1, 72-98.

Bock, R.D. & Wood, R. (1971). Test theory. *Annual Review of psychology*, 22, 193-224.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, 10. 291-299.

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing*, 7, 13-30.

Briere, E.J. , & Hinofotis, F. B. (Eds.). (1979). *Concepts in Language Testing: Some recent studies*. Washington, DC : TESOL.

Brooks, N. (1964). *Language and Language Learning*. New York: Harcourt Brace.

Brown, R. W. (1915). *How the French Boy Learn to Write*. Cambridge: Harvard University Press.

Brown, J.D. (1988). *Understanding Research in Second Language Learning*. Cambridge: Cambridge University Press.

_____. (1996). *Testing in Language Programs*. New Jersey: Prentice Hall Regents.

Buck, G. (1990). *The testing of second language listening comprehension*. Unpublished PhD Thesis. University of Lancaster.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing*, 8, 67-91.

_____ (1992). Translation as a language testing procedure: does it work? *Language Testing*, 9, 123-148.

Bullock, B. & Lantolf, J. P. (1987). Text gern and cloze testing in L2. *Rassegna Italiana di Linguistica Applicata*, 3. 95-104.

Carroll, B.J. (1980). *Testing communicative performance*. Oxford: Pergamon Press Ltd.

_____ (1985). Second language performance testing for university and profession contexts. In P.C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 73-78). Ottawa: University of Ottawa Press.

_____ and Hall, P.J. (1985). *Make Your own language tests*. Oxford: Pergamon Press Ltd.

Carrol, J. B. (1953). *The Study of Language: A Survey of Linguistics and Other Related Disciplines in America*. Cambridge, Mass.: Harvard University Press & London: Oxford University Press.

_____ (1961). Fundamental considerations in testing for English language proficiency of foreign students. In Center for Applied Linguistics, *Testing the English Proficiency of Foreign Students*. Washington, D. C.: Center for Applied Linguistics: 30-40.

_____ (1968). The psychology of language testing. In A. Davies, (Ed), *Language Testing Symposium*. London: Oxford University Press: 46-69.

_____ Carton, A. S. & Wilds, C. P. (1959). An investigation of cloze items in the measurement of achievement in foreign language. Cambridge, Mass.: Harvard Graduate School of Education. *ERIC ED* 021 513.

Cattell, R. B. (1964). Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology*, 55, 1-22.

Chapelle, C. A., & Abraham, R. A. (1990). Cloze method: what difference does it make? *Language Testing*, 7, 121-146.

Chihara, T., Oller, J. W. Weaver, K. A., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning*, 27, 63-73.

Chomsky, N. (1964). Current issues in linguistic theory. In J. A. Fodor and J. J. Katz (eds.). *The Structure of Language*. Englewood Cliffs: Prentice-Hall.

_____ (1965). Three models for the Description of language. In Luce, R.D., R. Bush, and E. Galanter, (eds.) 1965. *Handbook of Mathematical Psychology*. New York: John Wiley & Sons.

Clapham, C.M. (1992). The effect of academic discipline on reading test performance. Paper given at *the Language Testing Research Colloquium*, Princeton, NJ.

Clark, J. L. D. (1972). *Foreign language testing: Theory and practice*. Philadelphia, Pa.: Center for Curriculum Development, Inc.

_____ (1978). Psychometric consideration in language testing. In Spolsky (ed.), *Advances in Language testing: Series 2.* 1978: 15-30.

Clemans, W. V. (1979). Test Administration. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 188-201). Washington, DC: American Council on Education

Clifford, R.T. (1978). Reliability and validity of language aspects contributing to oral proficiency of prospective teachers of German. In J.L.D. Clark (Ed.), *Direct testing of speaking proficiency: Theory and application* (pp. 191-109). Princeton, NJ: Educational Testing Service.

Clifford, R.T. (1981). Convergent and discriminant validation of integrated and unitary language skills: The need for a research model. In L. Palmer & B. Spolsky (Eds.), *Papers in Language Testing*: 1967-74 (pp. 62-70). Washington, DC: TESOL.

Cohen, A. D. (1980). *Testing language ability in the classroom.* Massachusetts: Newbury House.

Cohen, A. (1994). *Assessing language ability in the classroom* (2nd ed.). Rowley, Mass: Newbury House/Heinle and Heinle.

Cooper, R.L. (1968). An elaborated language testing model. In J.A. Upshur & J. Fata (Eds), Problems in foreign language testing. *Language Learning Special Issue*. (No. 3, pp. 57-72). Ann Arbor, Mich., : Research Club in Language Learning.

Criper, C., & Davies, A. (1988). *ELTS validation project report*. London : The British Council and the University of Cambridge Local Examination Syndicate.

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Chicago, Ill.: Holt Rinehart Winston.

Cronbach, L. J. (1971). Validity in R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-597). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement*, 5, 99-108.

Cronbach, L. J. (1988). Construct validation after thirty years. In Robert L. Linn (Ed), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana, Ill. : University of Illinois Press.

Crowford, A. N. (1970). *The cloze procedure as a measure of the reading comprehension of the elementary level Mexican-American and Anglo-American children.* Unpublished doctoral dissertation. University of California, Los Angles.

Culhane, T., Klien-Braley, C., & Stevenson, D. K. (Eds.). (1981). *Practice and problems in language testing 4. Occasional Papers*, No. 26. Colchester, Essex: Department of Language and Linguistics, University of Essex.

Darnell, D. K. (1968). The development of an English language proficiency test of foreign students using a clozentrophy procedure. *ERIC ED* 024-039.

Dahan, H. (1996). Error analysis of Arabic writing skills (Malay version). *Journal of Educational Research*, Vol. 17. Kuala Lumpur: Faculty of Education, University of Malaya.

Davies, A. (1964). *English Proficiency Test Battery*, Version A. London: British Council.

_____ (1965). *Proficiency in English as a second language*. Unpublished Ph.D thesis. University of Birmingham.

_____ (ed.). (1970). *Language Testing Symposium: A Psycholinguistic Approach.* London: Oxford University Press.

_____ (1977). The construction of language tests. In J. P. B. Allen & A. Davies (Eds.), *Testing and experimental methods*. The Edinburgh Course in Applied Linguistics (Vol. 4, pp. 38-194). London: Oxford University Press.

_____ (1978). Language testing. In *Language Teaching and Linguistic Abstract* 11. 3/4: 145-160 and 215-232.

_____ (1984). Validating three tests of English Language proficiency . *Language Testing*, 1, 50-69.

_____ (1990). *Principles of language testing*. Oxford: Blackwell.

de Jong, J. H. A. L., & Glas, C. A. W. (1987). Validation of listening comprehension tests using item response theory. *Language Testing*, 4, 170-194.

Dornyei, Z. & Katona, L. (1992). Validation of the C-test amongst Hungarian ELL learners. *Language Testing*, 9, 187-206.

Douglas, D., & Selinker, L. (1985). Principles for language tests within the "discourse domains" theory of interlanguage: Research, test construction and interpretation. *Language Testing*, 2, 205-26.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd edition). Englewood Cliffs, NJ: Prentice Hall.

Ebel, R. L. & Frisbie, D.A. (1991). *Essentials of Educational Measurement*. (5th edition). Englewood Cliffs, NJ: Prentice-Hall.

Elder, C. (1995). The effect of language background on 'foreign language test performance: Problems of classification and measurement. *Language Testing Update*, 17, 36-38.

Estrada, F. X. (1969). *The effect of increasing syntactic complexity on reading comprehension*. Unpublished masters thesis. University of California, Los Angles.

ETS. (1989). *Understanding TOEFL*: Workbook. Princeton.

Fachrurrazy. (1989). Dictation as a device for testing English as a foreign language. *GUIDELINES: A Periodical for Classroom Language Teachers*. 11. 2: 48-60.

Farhady, H. (1980). *Justification, development, and validation of functional language tests*. Unpublished Ph.D. dissertation, University of California at Los Angeles.

Feldt, L. S., & R. L. Brennan. (1988). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education /Macmillan.

Finlayson, D. L. (1951). The reliability of the marking of essays, *British Journal of Educational Psychology*, 21/2: 126-34.

Fouly, K. and Cziko, A. (1985). Determining the reliability, validity, and scalability of the graduated dictation test. *Language Learning*, 35. 4 : 555-566.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, 10, 133-170.

Fries, C. C. (1945). *Teaching and Learning English as a Second Language*. Michigan: University Press.

Fulcher, G. (1994). Some priority areas for oral language testing. *Language Testing Update*, 15, 39-47.

Gaise, S. J., Gradman, H. L., and Spolsky, B. (1977). Toward the measurement of functional proficiency: Contextualization of the noise test. *TESOL Quarterly*, vol. 1, no. 1. pp. 51-57.

Al-Ghamdi, G. A. T. (1986). *English Proficiency in the Saudi Air Academy: Validating a*

new Test Battery. Unpublished Ph.D thesis. University of Edinburgh: Department of Linguistics.

Gouin, F. (1894). *The Arts of Teaching and Studying Languages*. trans. H. Swan & V. Betis. New York: Scribner.

Gradman, H. L. and Gaies, S. J. (1975). Reduced redundancy and error analysis: a study of selected performance on the "noise test". *Paper presented at the 4th AILA congress in Stuttgart*.

Green, R. & Weir, C. (1998). *Statistical Analysis for Language Testing and Evaluation*. Monograph: University of Reading.

Griffin, P. E. (1985). The use of latent trait models in the calibration of tests of spoken language in large-scale selection-placement programs. In Y. P. Lee, A.C. Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 149-161). Oxford: Pergamon Press.

Gronlund, N. E. (1982). *Constructing Achievement Tests*. (3rd ed.), New Jersey: Prentice-Hall.

_____ (1985). *Measurement and evaluation in teaching* (5th ed.). New York: Macmillan.

Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. *Language Testing*, 3, 159-85.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-81.

Guttman, L. (1953). Reliability formulas that do not assume experimental independence. *Psychometrika*, 18, 225-39.

Hair, J.F., Anderson, R. E., and Black, W. C. (1995). *Multivariate Data Analysis with Reading* (4th edition). NJ: Prentice-Hall (International Edition).

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). Summaries of studies involving the Test of English as a Foreign Language, 1963-1982. *TOEFL Research Report* 16. Princeton, NJ: Educational Testing Service.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing*, 6, 47-76.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199-230). Baltimore, Md.: The Johns Hopkins University Press.

Hambleton, R. K. (1988). Principles and selected applications of item response theory. Linn, R. L. (Ed.), (1988). *Educational measurement* (3rd ed., pp. 147-200). New York: American Council on Education/Macmillan.

Hanzeli, V. E. (1979). Cloze tests in French as a foreign language: Error analysis. In E. J. Briere and F. B. Hinofotis (eds.). Concepts in Language Testing: Some Recent Studies. Washington D. C. : *TESOL*, pp. 3-11.

Harris, D. F. (1969). *A Language Testing Handbook*. New York: McGraw Hill.

Harris, D. P. (1970). *The Linguistics of Language Testing*. In Davies, A. (ed). (1970): 36-45.

_____ (1988). *Testing English as a Second Language*. New York: McGraw-Hill Book Company.

Harrison, A. (1983). *A language testing handbook*. London: Macmillan Press.

Heaton, J. B. (1979). *Writing English Language Test*. (5th ed.). Longman.

Henning, G. (1987). *A guide to language testing*. Cambridge, Mass.: Newbury House.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, 9, 1-11.

Henrysson, S. (1971). Gathering, Analyzing, and Using Data on Test Items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130-159). Washington, DC: American Council on Education

House, E. R. (1977). *The logic of evaluative argument*. C. S. E. University of California monograph series in Education. no. 7.

Howell, T.M. (1975). *Survey of British based examinations for overseas students*. MA report. University of London: Birkback Colege.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6, 153-60.

Hughes, A. (1992). *Testing for language teacher*. (4th ed.). Cambridge University Press.

Ilyin, D. (1972). *Ilyin Oral Interview*. Rowley, Mass.: Newbury House.

Ingram, E. (1968). Attainment and diagnostic testing. In Davies, A. (ed)., *Language Testing symposium*, 1968: 70-97.

_____ (1977). Basic concepts in testing. In J. P. B. Allen & A. Davies (Eds.), *Testing and experimental methods*. The Edinburgh Course in Applied Linguistics (Vol. 4, pp. 11-37). London: Oxford University Press.

_____ (1978). The psycholinguistic basis. In Spolsky (ed.). *Advances in language testing: series 2*. 1978: 1-14.

Irvin, P., Atai, P., and Oller, J. W. (1974). Cloze, dictation, and the test of English as a foreign language. *Language Learning*, 24 : 245-252.

Jafarpur, A. (1987). The short-context technique: an laternative for testing reading

comprehension. *Language Testing*, 4, 195-220.

Johnson, K. T. (1980). Questioning some assumption about cloze testing. In J. A. Read (ed.). *Directions in Language Testing*. Singapore : SEAMEO Regional Language Centre. pp. 177-206.

Jones, R. L. (1979). The oral interview of the Foreign Service Institute. In B. Spolsky (Ed.), *Some major tests. Advances in language testing: series 1* (pp. 104-115). Arlington, Va: Center for Aplied Linguistics.

Jones, R. L. (1985). Some basic considerations in testing oral proficiency. In Y. P. Lee, A. C. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 77-84). Oxford: Pergamon Press.

Jonz, J. (1976). Improving the basic egg: The multiple-choice cloze. *Language Learning*, 26, 255-65.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, 6, 125-160.

Kattan, J. (1990). *The construction and validation of an EAP test for second year English and nursing majors at Bethlehem University*. Unpublished Ph.D thesis: University of Lancaster.

Kelly, L. G. (1969). *25 Centuries of Language Teaching*. MA: Newbury House.

Kerlinger, F. N. (1973). *Foundations of Behavioral Research*. New York: Holt, Rinehart, and Winston.

Kitao, S. K., & Kitao, K. (1996). Testing listening. [On-line]. *The Internet TESL Journal*, 2 (7). Available: http://www.aitech.ac.jp/~iteslj/Articles/Kitao-TestingListening.html.

Klein-Braley, C. (1983). A cloze is a cloze is a question. In Oller (ed.). *Issues in Language Testing*. Rowley, MA: Newbury House.

_____ (1985). A cloze-up on the c-test: A study in the construct validation of authentic tests. *Language Testing*, 2. 76-104.

Klien-Braley, C., & Stevenson, D. K. (Eds.). (1981). *Practice and problems in language testing 1*. Frankfurt: Verlag Peter D. Lang.

Kohenen, V., von Essen, H., & Klien-Braley, C. (Eds.). (1985). *Practice and problems in language testing 8*. Tampere, Finland: Finnish Association for Applied Linguistics.

Krzanowski, W. J., & Woods, A.J. (1984). Statistical aspects of reliability in language testing. *Language Testing*, 1, 1-20.

Lado, R. (1957). *Linguistics across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor: University of Michigan Press.

_____ (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. New York: McGraw-Hill.

_____ (1964). *Language Testing: A Scientific Approach*. New York: McGraw Hill.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, 19, 673-87.

Lee, Y.P. et al. (1985). *New directions in language testing.* Eds. Oxford: Pergamon Press.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. *Applied Psychological Measurement*, 4, 547-61.

Lombardo, M. (1981). The construction and validation of the listening and reading componenets of the English as a second language assessment battery. In R. V. Padilla (ed.) *Bilingual Education Technology*. Ypsilanti : Eastern Michigan University.

Low, G. D., & Lee, Y. P. (1985). How shall a test be referenced? In Y.P. Lee, A. C. Y. Fok, R. Lord, & G. Low (Eds.). *New directions in language testing* (pp. 119-126). Oxford: Pergamon Press.

Lumsden, J. (1976). Test theory. *Annual Review of psychology*. 27, 251-280.

MacGinitie, W. H. (1961). Contextual constraint in English prose paragraphs. *Journal of Psychology* 51. 1121-1130.

Madaus, G. F. (1983). *The courts, validity, and minimum competency testing*. Boston, Mass. : Kluwer-Nijhoff.

Magnusson, D. (1967). *Test Theory*. Reading, MA: Addison-Wesley.

Mathews, J. C. (1985). *Examinations: A Commentary*. London: George Allen and Unwin.

Mathews, T. J. (1992). *Development of a foreign language placement test using item response scoring on a multiple-choice cloze test*. Unpublished Ph.D dissertation. University of Deleware.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142-154.

_____, & Jones, G. (1988). Vocabulary size as a placement indicator. In *British Studies in applied linguistics*, 3. Nottingham, England.

Messick, S. A. (1980). Test validity and the ethics of assesment. *American Psychologist*, 35, 1012-27.

_____ (1988). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Mcmillan.

Mollenkopf, W.G. (1950). An experimental study of the effect on item analysis data of

changing item placement and test time limit. *Psychometrika*, 15, 291-316.

Moller, A. D. (1981). Assessing proficiency in English for use in further study. In J.A.S. Read (ed.). *Directions in Language Testing*. pp. 58-71. Singapore: Singapore University Press/ SEAMEO Regional Language Centre.

_____ (1982). *A study in the validation of proficiency tests of English as a foreign language*. Unpublished Ph.D thesis. University of Edinburgh. Department of Linguistics.

Morrow, K. (1979). Communicative Language Testing: Revolution or Evolution? In C. J. Brumfit and K. Johnson (eds.), *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press.

_____ (1986). The Evaluation of Tests of Communicative Performance. In M. Portal (ed.), *Innovations in Language Testing*. Windsor, Berks: NFER-Nelson.

Moser, C. A. & Kalton, G. (1971). *Survey Methods in Social Investigation*. 2nd. ed. London: Heinemann.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, 7, 191-205.

Moulton, W. G. (1961). Linguistics and language teaching in the United States, 1940-1960. In Mohrman, Christine, Sommerfelt, Alf., and Whatmough, Joshua (Eds.). *Trends in European and American Linguistics*, 1930-1960. Utrecht: Spectrum Publisher.

Muḥammad, M. A. (1989). *Language Testing* (Arabic version). Riyadh: University of King Saud Printing.

Mullen, K.A. (1979). More on close tests as tests of proficiency in English as a second language. In E.J. Briere and F.B. Hinofotis (Eds.), *Concepts in Language Testing: Some Recent Studies*, pp.21-32. Washington, D.C.: TESOL.

Noll, V. H., Scannell, D. P., and Craig, R. C. (1979). *Introduction to Educational Measurement*. 4th ed. Boston: Houghton Mifflin Company.

North, B. (1994). Item banker: A testing tool for Language teachers. *Language Testing Update*, 16, 85-97.

Oller, J. W. (1971). Dictation as a device for testing foreign-language proficiency. *English Language Teaching*, 25. 3: 254-259.

_____ (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *The Modern Language Journal* LVI. 3: 151-158.

_____ (1973). Cloze tests of second language proficiency and what they measure. *Language Learning* 23. 105-118.

_____ (1979). *Language Test at School : A Pragmatic Approach*. London: Longman.

_____ & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning* 21. 2: 183-191.

_____ & Perkins, K. (1969). *Research in language testing*. Massachusetts: Newbury House Publishers, Inc.

_____ & Perkins, K. (1969). *Language in Education: testing and tests*. Massachusetts: Newbury House Publishers, Inc.

_____ and Streiff, V. (1975). Dictation: A test of grammar-based expectancies. *English Language Teaching Journal*, 30. 1: 25-36.

Otter, H. S. (1968). *A Functional Language Examination*: The Modern Language Association Examination Project. London: Oxford University Press.

Pachinburavan, N. (1985). *Development and validation of an English placement test for freshman students at Khonkaen University*. Unpublished Ph.D dissertation. Washington State University. Department of Education.

Pack, A. C. (1973). Cloze testing and procedure. *TESL Reporter* 6. 1-2.

Palmer, A. S. (1981). Measurement of reliability and validity of two picture-description tests of oral communication. In L. Palmer & B. Spolsky (Eds.), *Papers in language testing: 1967-74* (pp.127-139). Washington, DC: TESOL.

Palmer, A. S., Groot, P. J. M., & Trosper, G. A. (Eds.). (1981). *The construct validation of tests of communicative competence*. Washington, DC: TESOL.

Perkins, K., Gupta, L., & Tammana, R. (1995). Predecting item difficulty in a reading comprehension test with an artificial neural network. *Language Testing*, 12, 34-53.

Peterson, C. R., & Cariter, F. A. (1975). Some theoritical problems and practical solutions in proficiency test validity. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp.105-13). Arlington, Va.: Center for Applied Linguistics.

Pilliner, A. E. G. (1968). Subjective and objective testing. In A. Davies (Ed.), *Language testing symposium. A psycholinguistic perspective*. London: Oxford University Press: 19-35.

Popham, W. J.(1967). *Educational Statistics: Use and Interpretation*. New York: Harper & Row Publishers.

_____, W. J. (1974). Selecting objectives and generating test items for objectives-based tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced evaluation.* CSE monograph series in evaluation, (No. 3, pp. 13-25). Los Angles, Calif.: Center for the Study of Evaluation, University of California at Los Angeles.

Portal, M. (1986). Methods of testing speaking in the Assesment of Performance Unit (APU) French surveys. In M. Portal (Ed.), *Innovations in language testing* (pp. 41-54). Windsor, Berks.: NFER-Nelson.

Raatz, U. (1985). Better theory for better tests? *Language Testing*, 2, 1, 61-75.

Rea, P.M. (1985). Language testing and the communicative language teaching curriculum. In Lee, Y.P. et al.. *New directions in language testing.* Eds. Oxford: Pergamon Press.

Richards, J., Platt, J., and Weber, H. (1985). *Longman Dictionary of Applied Linguistics*. London: Longman.

Rivers, W. (1968). *Teaching Foreign Language Skills*. Chicago: University of Chicago Press.

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing*, 9, 173-186.

Rowntree, D. (1991). *Statistics without tears, A Primer for non-mathematicians*. London: Penguin Books.

Ruddell, R. B. (1964). A study of the cloze comprehension technique in relation to structurally controlled reading material. *Proceedings of the International Reading Association* 9. 298-303.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assesment of writing*. New Jersey: Ablex.

Salvi, R. (1991). A communicative approach to testing written English in non native speakers. *Rassegna Italiana di Linguistica Applicata*, 23, 67-91.

Sarig, G. (1989). Testing meaning construction: can we do it fairly? *Language Testing*, 6, 77-94.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing*, 8, 95-111.

Savard, J-G. (1968). A proposed system for classifying language tests. In J. A. Upshur, & J. Fata (Eds.), *Problems in foreign language testing. Language Learning Special Issue* (No. 3, pp. 176-174). Ann Arbor, Mich.: Research Club in Language Learning.

Sax, G. (1979). *Foundation of Educational Research*. New Jersey: Prentice-Hall, Inc.

Scott, M. L. (1996). Examining validity in performance test: The listening summary translation exam (LSTE) - Spanish version. *Language Testing*, 13, 83-109.

Seliger, H. W. (1975). Two experiments in foreign language testing and acquisition. *Paper presented at the 4th AILA congress in Stuttgart.*

Shohamy, E. (1983). Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew. In J.W. Oller (ed.). *Issues in Language Testing Research*. Rowley, MA: Newbury House. pp 229-36.

_____ (1994). The validity of direct versus semi-direct oral tests. *Language Testing*,

11. 99-124.

_____ & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question types. *Language Testing*, 8, 1, 23-40.

Ṣīni, M. I. (1978). Designing a test for Arabic Language as a foreign language. In M. H. Bakalla (Eds.), *Proseedings of the First International Symposium on Teaching Arabic to Non-Arabic Speakers*. (Vol. 2, pp. 195-220). Riyadh.

Somaratne, W. (1957). *Aids and Tests in the Teaching of English*. London: Oxford University Press.

Spolsky, B. (1967). *Do they know enough English*? Some notes on the problems of assessing the proficiency in English of foreign students. In Wigglesworth (ed.). pp. 30-43.

_____ (1968). Language testing: The problem of validation. *TESOL Quarterly*, 2, 88-94.

_____ (1973). What does it mean to know a language; or how do you get someone to perform his competence? In J. W. Oller and J. C. Richards (eds.). *Focus on the Learner: Pragmatic Perspectives for the Language Teacher*. Massa: Newbury House Publishers, Inc.

_____ (ed.). (1978). *Advances in Language Testing: Series 2, Approaches to Language Testing*. Arlington, Virginia: Center for Applied Linguistics.

_____ (1995). The prehistory of TOEFL. *Language Testing*, 7, 98-118.

_____ Murphy, P., Holm, W., & Ferrel, A. (1972). Three functional tests of oral proficiency. *TESOL Quarterly*, 63, 221-235.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington, DC: American Council on Education.

Stansfield, C. W. (1985). A history of dictation in foreign language teaching and testing. *The Modern Language Journal*, 69. 2: 121-128.

Stevenson, D. K. (1985). Authenticity, validity and a tea party. *Language Testing*, 2: 41-7.

Stubbs, J.B. & Tucker, G.R. (1974). The cloze test as a measure of English proficiency. *Modern Language Journal*, 58, 5/6, 239-41.

Subkoviak, M. J. & Baker, F. B. (1977). *Review of Research in Education*, 5, 275-319.

Sweet, H. (1899). *The Practical Studies of Languages*. London: Dent.

Taylor, C. V. (1980). Dictation as a test of English proficiency. *RELC Journal*, 11. 2: 88-92.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30: 415-438.

_____ (1956). Recent developments in the use of the cloze procedure. *Journalism Quarterly*, 33 : 42-49.

Theunissen, T. J. J. M. (1987). Text banking and test design. *Language Testing*, 4, 1-8.

Thorndike, R. L. (1971) *Educational measurement* (2nd edition) (ed.). Washington DC: American Council on Education.

Tinkelman, S.N. (1971). Planning the objective test. In R. L. Thorndike (ed.), *Educational measurement* (2nd ed., pp. 46-80). Washington, DC: American Council on Education

Twiest, M. M. (1988). *Construction and validation of a test of basic process skills for the elementary and middle grades using different methods of test administration.* Unpublished EdD dissertation, University of Georgia.

Upshur, J. A. (1962). Language proficiency testing and the constrastive analysis dilemma. *Language Learning* 12: 123-128.

_____ (1972). Productive communication testing: A progress report. In Perren, G.E. and J.L.M. Trim (eds.). *Application of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics, Cambridge. 1969.* Cambridge: Cambridge University Press. 435-441.

Valdman, A. (Ed.). (1988). The assesment of foreign language oral proficiency. *Studies in Second Language Acquisition*, 10, 2.

Valette, R. M. (1964). The use of the dictée in the French language classroom. *Modern Language Journal*, 48 : 431-434.

_____ (1967). *Modern Language Testing: A Handbook.* New York: Harcout Brace, & World.

_____ (1977). *Modern Language Testing* (2nd. ed.). New York: Harcout Brace, Jovanovich.

Vaughan James, C. and S. Rouve. (1973). *Survey of Curricula and Performance in Modern Languages 1971-2.* London: Centre for Information on Language Testing and Research.

Vernon, P. E. and Milligan, G. D. (1954). A further study of the reliability of English essays, *British Journal of Statistical Psychology*, 7/2: 65-74.

Wainman, H. (1979). Cloze testing of second language learners. *English Language Teaching Journal*, 33. 2: 126-132.

Wall, D., C. Clapham, & J. C. Alderson. (1994). Evaluating a placement test. *Language Testing*, 11, 321-344.

Weir, C. (1990). *Communicative language testing.* New York: Prentice Hall International.

_____ (1993). *Understanding and developing language tests*. New York: Prentice Hall.

Weiss, D. J. & Davidson, M. L. (1981). Test theory and methods. *Annual Review of Psychology*, 32, 629-658.

Whiteson, V. (1972). The correlation of auditory comprehension with general language proficiency. *Audio-Visual Language Journal*, 10, 2, pp. 89-91.

Witt, R. de. (1992). *How to prepare for IELTS*. UK: The British Council.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, Ill.: MESA Press.

Texts sources:

Al-Qur'ān al Karīm.

Center for Applied Linguistics. (1997). *Arabic Proficiency Test* (APT). Washington DC: Center for Applied Linguistics.

King Sa`ūd University. (1985). *Arabic language testing for non-native speakers of Arabic*. Riyadh: King Sa`ūd University Press.

Shawkānī, M. Λ. (n.d.). *Nayl al-awṭār min aḥādith sayyid al-akhbār. Sharḥ muntaqā' al-akhbār*. Vol. I-ix. Beirut (1973).

Zuhaylī, W. (1984). *Al-Fiqh al-Islāmī wa adillatuh*. 3rd. Edtn. Damascus: Dār al-Fikr.

Appendix A Test papers and Questionnaires

A.1 Test papers at the AIS

A.1.1 Placement test (Pre-AIS) 1996/97

PRA AKADEMI PENGAJIAN ISLAM UNIVERSITI MALAYA

Ujian Penilaian Bahasa Arab Bagi Pelajar Baru

Sesi 1996/1997.

Tarikh : ..... Masa :

==========

Nama :.....

Kad Pengenalan: .....

الإمتحان انتنسبتى

فـي

اللغة العربيـــة

ملاحظة :

سود الحرف الذي يسبق الإجابة الصحيحة علي ورقة الإجابة بقلم رصاص .

A = أ

B = ب

C = ج

تحتوي هذه الورقة علي مائة سؤال ، لكل سؤال درجة واحدة فقط .

لا تفتح هذه الورقة إلا بعد صدور الأمر بذلك .

(١) الكلمات الآتية أسماء إلا :-

أ - اجتهاد
ب - الجامعة
ج - التحق

(٢) الجمل الآتية اسمية إلا :-

أ - الطلاب جالسون علي مقاعد الدرس .
ب - في الحديقة أزهار مفتحة .
ج - رأيت الهلال بين السحاب .

(٣) ما يأتي جمل فعلية إلا :-

أ - يذهب الطلاب إلي الجامعة .
ب - تقدم المرء موقوف علي حسن أخلاقه .
ج - إلتحق طالب جديد بأكاديمية إسلامية بجامعة ملايا .

(٤) الحركة الصحيحة في آخر " درس " في الجملة الآتية (علمني أبي درس قواعد اللغة العربية) هي :-

أ - الفتحة
ب - الكسرة
ج - الضمة

(٥) أصح جواب للسؤال الآتي : -

السؤال : (لماذا تتعلم اللغة العربية ؟) هو :

أ - أتعلم اللغة العربية في المدرسة العربية .
ب - أتعلم اللغة العربية لأفهم القرآن .
ج - تعلمي اللغة العربية كتعلمي اللغة الإنجليزية .

(٦ - ٨) قال الله تعالي : (من عمل صالحا من ذكر وأنثي وهو مؤمن فلنحيينه حياة طيبة)

(٦) لفظ صالحا في الآية يعرب :-

أ - صفة
ب - حالا
ج - مفعولا به

(٧) قوله " وهو مؤمن " في الآية يعرب :-

أ - حالا
ب - مفعولا
ج - مفعولا مطلقا

(٨) كلمة " طيبة " في الآية يعرب :-

أ - مضافا اليه
ب - خبرا
ج - صفة

(٩) الجمل الآتية صحيحة إلا :-

أ - نحن طلا ب الجامعة
ب - إنما الحيا ة جهاد
ج - الجملة التي اشتريتها أمس فائدة جدا

(١٠) أكمل الجملة الآتية بإحدي الكلمات الآتية :
((أكل الأطفال حتي ..... ثم أخذ يضحكون ويلعبون حتي ناموا))

أ - يشبعو ن
ب - يشبع
ج - شبعوا

(١١) املأ الفراغ بإحدي الكلمات الآتية :-

(وصل الطالبا ن .....)
أ - الجديد مبكرا
ب - الجديدان مبكران
ج - الجديدا ن مبكرين

(١٢) أصح ترجمة لقول بعضهم (ساي ادا سبواه بوكو دان ساي تله باج بوكو ايت سبايق تيڬ كالي) هي :-

أ - يوجد عندي واحد كتاب وقرأ ت ذلك كثير ثلاث مرات
ب - أ نا موجود كتاب واحد وقرأ ت ذلك الكتاب ثلا ث مرة
ج - لي كتاب وقد قرأت الكتاب ثلاث مرات

(١٣) أصح ترجمة لقول بعضنا (ساي ادا سئورغ كاون يغ سوك ماكن دوريان هي :-

أ - انا يوجد صديق الذي يحب يأكل الدوريان
ب - لي صديق يحب أكل الدوريان
ج - عندي صديق الذي يحب يأكل الدوريان

(١٤) أنسب حرف للفراغ الآتي (لا بد ..... المسلم أن يصوم رمضان) هو : -

أ - من
ب - علي
ج - إلى ~~الام~~

(١٥) أنسب فعل للفراغ الآتي (..... المسلمون الي مكة) هو : -

أ - سافر
ب - سافروا
ج - سافرت

(١٦) ما معني الكلمة التي تحتها خط فيما يأتي :
(ارتدي زيد ملابسه ليذهب إلي المسجد)

أ - كوي
ب - غسل
ج - لبس

(١٧) أصح كلام يقال لمن لا يعرف نجاح زيد في الإمتحان :

أ - إ ن زيد ا ناجح في الامتحان
ب - زيد ناجح في الامتحان
ج - والله إن زيدا ناجح في الامتحان

(١٨) أصح ترجمة للعبارة الآتية : (زيد طالب زكي) هي :

أ - زيد فلاجر يغ فاليغ جرديق
ب - فلاجر يغ جرديق ايت ايالة زيد
ج - زيد ايالة سئورغ فلاجر يغ جرديق

(١٩) عين من الجمل الآتية جملة فيها فعل ناقص :

أ - جئت إليك وأنا مطمئن البال
ب - جاء زيد ماشيا مبتسما
ج - أمست السماء أن تمطر

(٢٠) الأسماء الآتية مؤنثة إلا :

أ - النار والجنة
ب - السماء والأرض
ج - النجم والقمر

(٢١) قال الله تعالي : (فيها سرر مرفوعة) فلفظ " سرر " مرفوع " .

أ - لأنه فاعل
ب - لأنه مبتدأ
ج - لأنه خبر

(٢٢) قال الله تعالي : ((إنما المؤمنون إخوة)) فكلمة " المؤمنون " :-

أ - اسم إن
ب - مبتدأ
ج - خبر

(٢٣) " القاضي في المحكمة " فكلمة " القاضي " هي :-

أ - اسم منقوص
ب - اسم مقصور
ج - اسم ممدود

(٢٤) " ذهب مصطفي إلي المستشفي " فكلمة مصطفي هي :

أ - اسم منقوص
ب - اسم مقصور
ج - اسم ممدود

(٢٥) عين الجملة الصحيحة من الجمل الآتية :

أ - ما زال فاطمة مجتهدة
ب - زالت فاطمة مجتهدة
ج - ما زالت فاطمة مجتهدة

(٢٦) عين من الجمل الآتية : " جملة فعلية فيها فعل ماض ناقص وخبرها مضارع " :-

أ - إن مع العسر يسرا
ب - بات الطفل يبكي
ج - قرأ التلميذ الدرس

(٢٧) عين جملة فعلية من الجمل الآتية " الفاعل فيها اسم منقوص " .

أ - حكم القاضي علي المتهم بالسجن المؤيد
ب - ضرب الأفعي بالعصي
ج - متي ينزل المطر يضم الزرع

(٢٨) عين جملة المبتدأ فيها نكرة وخبرها شبه جملة :-

أ - الطفلة في الحجرة
ب - إمرأة جميلة في الحجرة
ج - في الفصل زيد .

(٢٩) عين جملة اسمية الخبر فيها مفرد :

أ - زيد أبوه مدرس
ب - المعلمون عالمون
ج - التلميذ ينام في الفصل

(٣٠) عين جملة اسمية المبتدأ فيها اسم الإشارة :-

أ - الذي ينام في الفصل كسلان
ب - هم ناجحون
ج - هؤلاء التلاميذ ناجحون

(٣١) " زيد المجتهد ناجح في الإمتحان " يعرب لفظ " المجتهد " :-

أ - صفة
ب - خبرا
ج - بدلا

عين جملة فيها " الممنوع من الصرف " :-

(٣٢) أ - قرأت كتاب قواعد اللغة العربية
ب - يصوم المسلمون في شهر رمضان
ج - تعلمت في أفضل المدارس

(٣٣) أ - آمنت بالإسلام دينا وبمحمد نبيا
ب - وبشرناه بإسحق نبيا من الصالحين
ج - وما محمد إلا رسول

(٣٤) أ - لا تأكل وأنت شبعان
ب - هل تحرص علي مرضاة أساتذتك
ج - هل تواظب علي الصلوات النافلة .

(٣٥) " زيد لهو المجتهد [فكلمة " المجتهد " تعرب :-

أ - صفة
ب - خبرا
ج - بدلا

(٣٦) عين المفعول به في العبارة الآتية ((احترم والديك أباك وأمك))

أ - والديك
ب - أباك
ج - أمك

(٣٧) قال الله تعالي ((وسيجزي الله الشاكرين)) المفعول به في الجملة هو :

أ - سيجزي
ب - الله
ج - الشاكرين

(٣٨) لفظ الجلالة في الجمل السابقة يعرب :-

أ - مبتدأ
ب - اسم فاعل
ج - فاعلا

(٣٩) الجملة السابقة عبارة عن :

أ - جملة فعلية
ب - جملة اسمية
ج - جملة حالية

(٤٠) " ليس النجاح سهلا " كلمة " النجاح " تعرب :-

أ - خبر ليس
ب - اسم ليس
ج - خبرا مقدما

(٤١) " كأن العلم نور " ؛ أي إجابة صحيحة ؟

أ - كأنْ من الأفعال الناقصة
ب - كأنْ من الأحرف المشبهة بالفعل
ج - كأنْ من الأسماء المبنية

(٤٢) لفظ " كأن " من الجملة السابقة تستعمل :-

أ - للتشبيه الأكيد
ب - للاستدراك
ج - المترجي

(٤٣) لفظ العلم في الجملة السابقة يعرب :

أ - خبر كأن مقدم
ب - اسم كأن مرفوع بالضمة الظاهرة .
ج - اسم كأن منصوب بالفتحة الظاهرة

(٤٤) " كأنك فاهم " لفظ كأن هنا يدل علي :-

أ - الاستدراك
ب - التشبيه
ج - الشك

(٤٥) أكمل الجملة الآتية بالكلمات المناسبة :-

(..... تلميذ ..... غاية الذكاء و .... الآن لدخول الامتحان)

أ - اختك - ذكي - يستعد
ب - أخوك - ذكي - استعد
ج - أخوك - ذكي - يستعد

(٤٦) إذا أردت أن تسأل عن عدد الأقلام التي اشتراها زيد تقول :

أ - بكم اشتريت هذه الأقلام يا زيد
ب - كم قلما اشتريت يا زيد
ج - كم أقلام اشتريت يا زيد

(٤٧) المدارسُ جمع من :

أ - المدرس
ب - الدار س
ج - المدرسة

(٤٨) إخوة جمع :

أ - أخت
ب - أخ
ج - أخوات

(٤٩) معني " عم "

أ - أخ الأم
ب - أخ الأب
ج - إبن العم

(٥٠) هات جمعا لكلمة " الأديب "

أ - الآدا ب
ب - الأديبة
ج - الأدباء

(٥١) يقول الأعمي : " يا رجلا خذ بيدي " لفظ رجلا " يعرب : -

أ - منادي مضاف
ب - منادي نكرة مقصودة مفرد ة
ج - منادي نكرة غير مقصورة

- قال الله تعالي : ((يا أيتها النفس المطمئنة ارجعي إلي ربك راضية مرضية فادخلي في عبادي وادخلي جنتي))

(٥٢) " يا "

أ - أداة النداء للقريب فقط
ب - أداة النداء للبعيد فقط
ج - أدا ة النداء للقريب والبعيد معا

(٥٣) " أية "

أ - منادي مرفوع بالضمة
ب - منادي منصوب بالفتحة
ج - منادي مبني علي الضم

(٥٤) " النفس "

أ - منادي
ب - بدل
ج - صفة

(٥٥) " ارجعي "

أ - فعل ماض
ب - فعل أمر
ج - فعل مضارع

(٥٦) " جنتي "

أ - صفة وموصوف
ب - مضاف ومضاف إليه مرفوع
ج - مفعول به وهو مضاف ومضاف إليه .
" اللهم شكرا "

(٥٧) لفظ "اللهم" : - شكرا .

أ - منادي مضاف
ب - منادي مفرد
ج - منادي شبيه بالمضا ف

(٥٨) لفظ (شكرا) في الجملة السابقة :

أ - حال منصوب
ب - مفعول به
ج - مفعول مطلق

(٥٩) أ ثا ث البيت معناه :

أ - زينة البيت
ب - أهل البيت
ج - الوالدا ن

(٦٠ - ٦٢) الجمل الآتية فيها الممنوع من الصرف إلا :

(٦٠) أ - صليت في مساجد كثيرة
ب - تعلمت في أفضل المدا رس
ج - سافرت إلي نيلم فوري

(٦١) أ - سمعت حديث يزيد
ب - فرشت غرفتي بسجاجيد جميلة
ج - بات النجم لا معا

(٦٢) أ - سمعت أقاصيص مختلفة
ب - ا ستمعت إلي أقاصيص مختلفة
ج - سمعت الأقاصيص المختلفة

(٦٣) " متي الإمتحان ؟ " " متي "

أ - مبتدأ مرفوع بالضمة مقدرة
ب - اسم الاستفهام في محل رفع خبر مقدم
ج - حرف الاستفهام

(٦٤) قال الله تعالي : ((يا أبت إني رأيت أحد عشر كوكبا))
" كوكبا " يعرب : -

أ - مفعولا به منصوب
ب - تمييزا منصوبا
ج - حالا منصوبا

(٦٥) " أحد عشر " في الجملة السابقة :-

أ - حال
ب - تمييز
ج - مفعول به

(٦٦) الجمل الآتية خطأ إلا : -

أ - عندي خمسة كتاب وسبعة قلم .
ب - اشتريت خمسة كتب وسبعة اقلام
ج - عندي واحد كتاب وثلاثة أقلام .

(٦٧) الجمل الآتية صحيحة إلا : -

أ - اشتعل الرأس شيبا
ب - وفجرنا الأرض عيونا
ج - وأختار موسي قومه سبعين رجالا

(٦٨) " العقلاء يعتمدون علي أنفسهم " لفظ " يعتمدون " : -

أ - فعل مضارع مرفوع بالضمة
ب - فعل مضارع مرفوع بالواو
ج - فعل مضارع مرفوع بثبوت النون

(٦٩) قال الله تعالي : ((لن تنالوا البرّ حتي تنفقوا مما تحبون)) لفظ " تنالوا":

أ - فعل مضارع مرفوع بالواو
ب - فعل ماض مبني علي السكون
ج - فعل مضارع منصوب بحذف النون

(٧٠) كلمة " البر " : -

أ - مضاف إليه
ب - مفعول به
ج - مفعول فيه

(٧١) " يا طالب العلم : لا تهمل واجبك " ، فكلمة " طالب " يعرب :

أ - منادي مفرد علم مبني علي الفتح
ب - منادي مضاف مبني علي الفتح
ج - منادي مضاف منصوب بالفتحة

(٧٢) عين من الجمل الآتية " جملة فعلية فعلها مضارع منفي "

أ - ينجح المجتهدون في الامتحان
ب - محمد لم يشرب اللبن
ج - لن يأكل زيد سمكا

(٧٣) الجمل الآتية " فعلها يدل علي قرب وقوع الخبر " إلا :

أ - أوشك الوقت أن ينتهي
ب - كانت الشمس تشرق
ج - لعل الحبيب قادم

(٧٤) " ظننت الامتحان سهلا " (الامتحان): -

أ - اسم ظن مرفوع بالضم
ب - خبر ظن مرفوع بالضم
ج - مفعول أول منصوب بالفتحة

(٧٥) " ظننت " هو :

أ - فعل ماض منصوب بالفتحة والتاء ؛ تاء التأنيث في محل رفع فاعل .

ب - فعل ماض ناقص مبني على الفتح والتاء ضمير متحرك في محل رفع فاعل .

ج - فعل ماض مبني على السكون ، والتاء في محل رفع فاعل .

(٧٦) " سهلا " : -

أ - مفعول ثان منصوب بالفتحة

ب - حال منصوب بالفتحة

ج - مفعول ثان منصوب بالألف .

(٧٧) " حفظت الدرس حفظا " " حفظا " : -

أ - مفعول به منصوب

ب - مفعول مطلق منصوب

ج - حال منصوب

(٧٨) " صفق المستمعون إعجابا " (إعجابا) يعرب : -

أ - مفعولا به

ب - مفعولا لأجله

ج - مفعولا مطلقا

(٧٩) " صليت الظهر أربع ركعات أداء لله تعالي : -
لفظ " أربع "

أ - مفعول لأجله

ب - تمييز

ج - نائب عن المفعول المطلق

(٨٠) كلمة " ركعات " في الجملة السابقة : -

أ - مفعول مطلق منصوب

ب - تمييز منصوب

ج - مفعول أجله

(٨١) " صمت شهر رمضان " : " شهر " يعرب : -

أ - مفعولا به
ب - مفعولا لأجله
ج - مفعولا فيه

(٨٢) عين جملة من الجمل الآتية : " المبتدأ فيها مرفوع بالواو وخبرها شبه جملة " : -

أ - أخوك نائم في الفصل
ب - أخوك شاهد التلفزيون
ج - أبوك في البيت

(٨٣) عمل كان وأخواتها : -

أ - ترفع المبتدأ اسما لها وتنصب الخبر خبرا لها
ب - تنصب المبتدأ اسما لها وترفع الخبر
ج - تنصب المبتدأ والخبر معا علي أنهما مفعوليه

(٨٤) من أخوات كان هي : -

أ - أصبح - ظل - بات
ب - أضحي - ليس - لعل
ج - صار - أمسى - كاد

(٨٥) عمل إ ن وأخواتها : -

أ - ترفع المبتدأ وتنصب الخبر
ب - تنصب المبتدأ والخبر معا
ج - تنصب المبتدأ وترفع الخبر

(٨٦) من أخوات إن هي :

أ - إن - لكن - لات
ب - ما - كان - أن
ج - لعل - لكن - ليت

(٨٧) " معني " لكن " : -

أ - للتشبيه
ب - للتمني
ج - الاستدراك

(٨٨) " كأنك شمس والملوك كواكب " و " شمس " تعرب : -

أ - اسم كأن مرفوع بالضم
ب - خبر كأن مرفوع بالضم
ج - فاعلا مرفوعا بالضم

(٨٩) من أدوات الاستثناء هي :-

أ - إلا - غير - سوى
ب - ليس - لا يكون - أو شك
ج - خلا - عدا - بات

(٩٠) " قرأت الكتاب إلا صفحة " " الكتاب " يعرب : -

أ - المستثني
ب - المستثني منه
ج - أداة الاستثناء

(٩١) نجح الطلاب إلا زيدا " زيدا " يعرب : -

أ - مستثني بإلا منصوب بالفتحة .
ب - فاعل مرفوع بالضمة
ج - مستثني بإلا مرفوع بالضمة

(٩٢) عين " المستثني " فيما يأتي : " صمت رمضان إلا يومين " : -

أ - رمضان
ب - صمت
ج - يومين

(٩٣) عين المستثني منه فيما يأتي : " قابلت أصدقائي إلا واحدا : -

أ - قابلت
ب - أصدقائي
ج - واحدا

(٩٤) " ما فاز اللاعبون إلا لاعبا " :

أ - يجب نصب المستثني .
ب - يجوز نصب المستثني
ج - يجب رفع المستثني

(٩٥) " وما محمد إلا رسول " لفظ " رسول " يعرب :-

أ - مستثني بإلا منصوب
ب - خبرا مرفوعا
ج - مبتدأ مرفوعا

(٩٦) " جاء الطلاب خلا زيدا " : زيد " يعرب

أ - مستثني منه
ب - مستثني
ج - مفعولا به

(٩٧) " قرأت الكتاب غير صفحة " كلمة " " غير " : -

أ - مستثني منصوب بالفتحة
ب - مفعول به منصوب بالفتحة
ج - مفعول فيه منصوب بالفتحة

(٩٨) كلمة " صفحة " في الجملة السابقة :-

أ - مستثني منصوب بالفتحة
ب - مضاف اليه مجرور بالكسرة
ج - مستثني مجرور بالكسرة

(٩٩) " لا كتاب في الحقيبة [redacted] " لفظ " كتاب "

أ - اسم " لا " المشبهة بليس منصوب
ب - اسم " لا " النافية للجنس مبني علي الفتح
ج - اسم " لا " النافية للجنس منصوب بالفتحة

(١٠٠) " ما شارع مزدحما " لفظ " مزدحما : -

أ - مفعول ثان منصوب بالفتحة
ب - خبر " ما " المشبهة بليس منصوب بالفتح
ج - خبر ما النافية للجنس منصوب بالألف .

A.1.2 Placement test 1995/96

الأكاديمية الاسلامية

بجامعة مالايا

اختبار تحديد المستوى

يوليو ١٩٩٥

الزمن : ساعة واحدة

الملحوظة :

١ـ أجب عن كل الأسئلة

٢ـ أن تكون الاجابة على ورقة الأسئلة ذاتها

تحتوى الأسئلة على خمس (٥) صفحات مطبوعة

الاسم : _____

الكلية : _____

رقم البطاقة الخصية : _____

السـؤال الأول

املأ الفراغ فى الجمل الآتية بالكلمات التى فى المربع الاتى :

هو ـ الانطلاق ـ من ـ اخلال ـ تتمتع

١ـ ان عمل المرأة خارج البيت يحررها ـــــ سلطة الرجل

٢ـ ان الغاء وظيفة البيت أو تسليمها لغير الزوجة ـــــ بنظام اجتماعى محترم .

٣ـ الحرية معناها ـــــ بلا قيود .

٤ـ الشهوة ـــــ بالطغيان وقوة الغلبة .

٥ـ كان الايمان بالله والعمل بشرعه ـــــ السلاح الذى يستعين به الانسان .

السـؤال الثانى

رتب الجمل الآتية ترتيبا صحيحا :ـ

١ـ كانوا / عباد أوثان / قبل الاسلام / العرب

٢ـ لحكومة / ما / لا قوة / بقوة شعبها / الا

٣ـ اصلاح الضمير / عليه / يترتب / اصلاح العقيدة

٤ـ ان أبطال / على وعى تام / الانتفاضة / بقضيتهم

٥ـ أحب اللـــه تعالى / من / أحب رســوله محمد

الســـــؤال الثالث

صحح الأخطاء الظاهرة فى الجمل الآتية :ـ

١ـ الداعية المسلم طبيبة و صيدلية يكشف الداء

٢ـ أن الغرب قد ترك الأخلاق الحميد للسيد المسيح

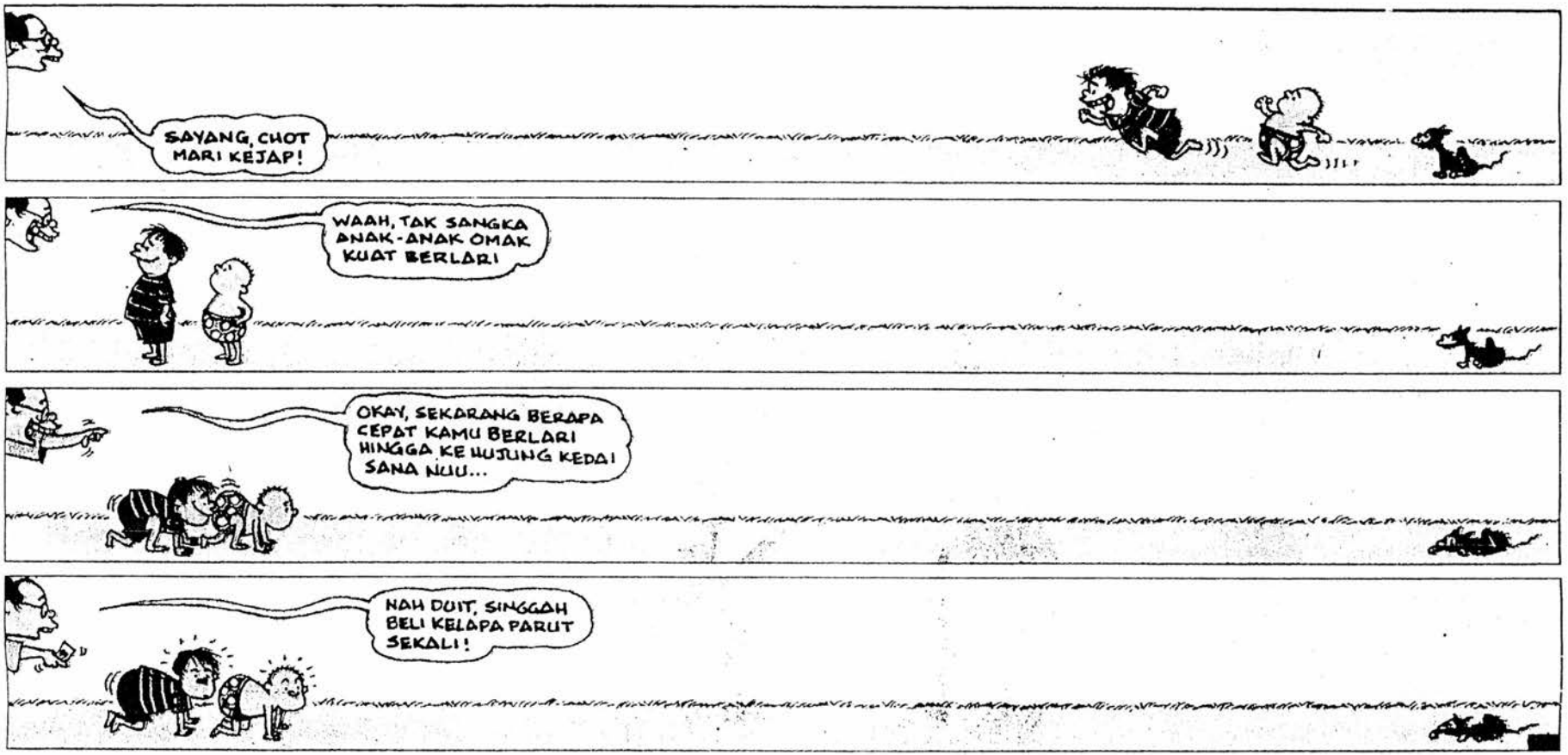
٣ـ من أهدافنا تخريج الطلبة المثقفون

٤ـ يجب على المتعلم أن تقبل على العلم بنية صافية نقية

٥ـ كل حضارة تعتمد لضمان بقائه واستمراره على تعليم أبنائها بالمخاطر الخارجية التى تهددها

السـؤال الرابع

اكتب فقرة قصيرة تعبر عما فهمت من الكاريكاتور الاتى :



A.1.3 Placement test 19996/97

AKADEMI PENGAJIAN ISLAM
UNIVERSITI MALAYA

UJIAN PENEMPATAN BAHASA ARAB
SESI 1996/97
JUN 1996
MASA : 1 JAM

ARAHAN : (a) JAWAB SEMUA SOALAN.
(b) JAWAPAN HENDAKLAH DITULIS DI DALAM KERTAS SOALAN INI.

NAMA : _____

NO. KAD PENGENALAN : _____

FAKULTI : _____

(Kertas ini mengandungi 5 soalan di dalam 6 halaman yang bercetak)

بسم الله الرحمن الرحيم

السؤال الأول

عيّن الكلمة المناسبة للفراغات الآتية :

جاء الإسلام فحرّر العقل (١) الخرافات ، ودفع المسلم إلى (٢) ينظر ماذا في السموات (٣) الأرض وعلّم المؤمنين أنّ (٤) ما في الكون مسخّر لـ (٥) .

وشجّع الإسلام على العلم و(٦) أنّ التعمّق في دراسة المخلوقات (٧) للتعرّف على الخالق و(٨) المسلمون القادرون لذلك التوجيه (٩) فلم يقتصروا على علوم (١٠) بل كان منهم علماء في الفلك والطبّ والصيدلية وغيرها من العلوم .

(١) أ . مِن	(٢) أ . سوف	(٣) أ . كـ	(٤) أ . نصف
ب . نحو	ب . أنْ	ب . فـ	ب . معظم
ج . عن	ج . كي	ج . و	ج . بعض
د . على	د . لن	د . لـ	د . جميع

(٥) أ . ـهم
ب . ـه
ج . ـها
د . ـهنّ

(٦) أ . أمر
ب . بيّن
ج . نهى
د . فتح

(٧) أ . مجموعة
ب . صالحة
ج . وسيلة
د . صعبة

(٨) أ . أجاب
ب . جاوب
ج . استجوب
د . استجاب

(٩) أ . الكريم
ب . القبيح
ج . البليد
د . المتخلّف

(١٠) أ . دين
ب . دينيّة
ج . دينيّ
د . الدينيّة

(١٠ درجات)

السؤال الثاني

حوّل الأفعال في الجمل الآتية من مبني للمعلوم إلى مبني للمجهول أو بالعكس وأعدْ كتابة الجملة :

(١) جَهَّزَ أبو بكرٍ الجيش لمحاربة المرتدّين .

_____ .

(٢) يُزْرَعُ بالمدينة كثيرٌ من الفواكه .

_____ .

(٣) يَزُورُ السائحون المدن التاريخيّة في ماليزيا .

_____ .

(٤) تُقْرَأُ الكتب والمجلاّت في تلك المكتبة .

_____ .

(٥) يَسْأَلُ المدير العاملات عن أعمالهنّ .

_____ .

(١٠ درجات)

السؤال الثالث

املأ الفراغات في الجمل الآتية بكلمةٍ مشتقّةٍ من المادة الّتي بين القوسين :

(١) كتاب الفقه الإسلامي _____ في هذا المعرض . (ع / ر / ض)

(٢) يجاهد الجنود في _____ عن الوطن . (د / ف / ع)

(٣) _____ المحاربون سهامهم نحو الأعداء . (ط / ل / ق)

(٤) افتتح السيّد الرئيس الاجتماع السنويّ بـ_____ الخطبة . (ل / ق / ي)

(٥) الوقاية خير من _____ . (ع / ل / ج)

(١٠ درجات)

السؤال الرابع

أكمل الجمل الآتية بالكلمة المناسبة من القائمة :

يحقّق - يجهل - دعا - ينقذ - يخشى

(١) _____ العميد جميع المحاضرين إلى الاجتماع اليوم .

(٢) استطاع الكشّاف أن _____ الولد من الغرق .

(٣) كان أخي _____ كلّ شيء عن هذه المدينة عندما حضر لأوّل مرّة .

(٤) _____ محمّد أنْ تتأخّر والدته في المطار .

(٥) استطاع فريقنا أن _____ نتيجةً ممتازةً .

(١٠ درجات)

السؤال الخامس

اختر الكلمة الغريبة في كلّ مجموعةٍ من المجموعات الآتية :

(١) جدّة - عمّة - خالة - زبدة - حفدة .

(٢) رسالة - بذلة - جريدة - صورة - مجلّة .

(٣) حذاء - موز - عنب - بيض - لبن .

(٤) كرافتة - جلابيّة - صدريّة - جبهة - قلنسوة .

(٥) كتف - فخذ - ريف - قدم - أنف .

(١٠ درجات)

م ز ف / مايو ١٩٩٦ ******** مَعَ تَمَنِّيَاتِنَا بِالنَّجَاحِ وَالتَّوْفِيقِ ********

A.1.4 Achievement test 1995/96

UNIVERSITI MALAYA

PEPEREKSAAN IJAZAH SARJANA MUDA SYARIAH

PEPEREKSAAN IJAZAH SARJANA MUDA USULUDDIN

TAHUN PERTAMA 1995/96

IB101 : BAHASA ARAB

MAC/APRIL 1996 MASA 3 JAM

==========

ملحوظة :

١ - أجب عن الأسئلة كلها.

٢ - الإجابة عن الأسئلة الموضوعية لا بد أن تكون على ورقة الإجابة الخاصة للحاسب الآلي .

(KERTAS SOALAN INI MENGANDUNGI 3 BAHAGIAN DALAM 13 HALAMAN YANG BERCETAK)

بسم الله الرحمن الرحيم

القسم الأول : العلوم العربية

المجموعة الأولى : النحو

(١٥ حرجة)

١ - مثل لما يأتي في جمل مفيدة :

١ - جملة فعلية : الفعل فيها يعرب بالحركة المقدرة للتعذر و الفاعل يعرب بالحركة المقدرة للثقل .

٢ - جملة مشتملة على المركب من ظرف المكان المبني على فتح الجزأين .

٣ - جملة اسمية الخبر فيها الاسم المنقوص المضاف إلى ياء المتكلم و كان مفرداً.

(٣ درجات)

ب - أجب عن الأسئلة الآتية :

١ - اذكر علامة إعراب الاسم المقصور و الفعل المضارع المعتل الآخر بالألف فى حالة النصب مع الإتيان بالأمثلة .

٢ - ما هو المعرب من الأفعال ؟ ، وضح ذلك مستدلا بالأمثلة . (٤ درجات)

ج - بين ضمير الرفع و النصب و الجر مستترا كان أو بارزا فى الكلمات التي تحتها خط :

١ - قال تعالى (قال رب إني ظلمت نفسي فاغفرلي ، فغفرله إنه هو التواب الرحيم)

٢ - قال تعالى (ربنا إننا سمعنا مناديا ينادي للإيمان ..)

٣ - قال تعالى (ربنا لا تزغ قلوبنا بعد إذ هديتنا)

(٣ درجات)

المجموعة الثانية : الصرف

أجب عن سؤالين فقط من الأسئلة الآتية :

ا - جرد الأفعال الآتية من الحروف الزائدة :

أخرج - استغفر - تصنّع - فهّم - انتقل

ب - بين نوع كل من الأسماء الجامدة التي تحتها خط فيما يأتي :

١ - هذا مكتب

٢ - من يزرع يحصد

٣ - لا تأخذ ما فى الدرج

٤ - أي كتاب تريد ؟

٥ - الكتاب الذي عندي جديد.

ج - عين الجامد و المشتق من الأسماء التالية :-

محمد - العلم - المسجد - كريم - البيت

(٥ درجات)

القسم الثاني المهارة اللغوية

المجموعة الأولى : القطعة و الترجمة

(٢٥ درجة)

اقرأ القطعة الآتية بتأنٍ ثم أجب عن الأسئلة التي تليها :

الإسراء رحلة قام بها النبي صلى الله عليه و سلم من المسجد الحرام إلى المسجد الأقصى، و المعراج رحلة سماوية ، قام بها صلى الله عليه وسلم من المسجد الأقصى إلى السموات العلى .....

الرحلتان كانتا فى ليلة واحدة ، و باشرهما الرسول الكريم كإنسان كامل . و قد أشار القرآن إلى القصة فى سورة الإسراء. و بعض الناس يقف عند القصة وقفة التأمل أو تردد، فيسأل : هل تتفق القصة و نوامس الله فى خلقه، فيكون هناك إنسان خلق من لحم ودم و يحتاج لكل مقومات المادية ، ثم يصعد إلى السموات ، مع أنا نعلم تخلخل الهواء فى مكان معلوم ، و فقدان الأوكسجين فى نقطة معلومة؟... كنا نقول لهؤلاء ... هى قدرة الله تعالى ، وسعت كل شئ فهو أمر ممكن لا يستحيل على قدرة الله تعالى ... و لكن ، هل أحطتم بكل العلم شارده ووارده ؟ الواقع أيها الإخوان أن العلم الحديث قد كشف تعليل ذلك فى أن الإنسان فيه عنصر آخر غير عناصر المادة ، ذلك هو العنصر النفساني ، الذي يطلق عليه عالم الروح و النفس . ولو كان العلم لم يصل بعد لحقيقته ، فإنه قد وصل إلى أن الروح لها من القدرة على الجسم ما تستطيع به أن تؤثر عليه و تحتجزه فتخضعه لقوانينها ، لا لقوانين المادة . و الواقع أن بعض الحوادث تفسرلنا ذلك ، فهناك بعض من صوفية الهند يستطيع أن يتحكم فى جسمه بقوة روحه

و يمكث أسبوعا ، و عندنا التنويم المغناطيسي الذي يجعل الروح تسيطر على الجسم ، فإذا هو عيون ترى ...

فالذي حدث فى قصة الإسراء و المعراج أن الله تعالى أفاض على نبيه الكريم قوة روحية عظمى ، تحكمت فى جسمه و سيطرت عليه ، و ليس معنى هذا أنه أسرى بالجسم دون الروح ، و إنما أسرى بالروح و الجسم ... وبعض الناس يتساءلون : ما حكمة الإسراء و المعراج ؟...

و أعتقد أن الإسراء و المعراج مادة أساسية فى منهاج التربية الإلهية ، ذلك أن الله تعالى أعد رسوله الكريم ليكون سيد المرسلين ، فلا بد أن يكون بمنزلة من العلم تفوق أي منزلة سواها من منازل البشر ، و لهذا طاف الله به فى السموات ليكون إيمانه رؤية و مشاهدة ، و ليس إيمانا نظريا ...

و هناك حكمة أخرى ، فيها سمو القدر و جلال المنزلة ، فالحق تبارك و تعالى قد فرض الصلاة على المسلمين ليلة الإسراء و المعراج ، و لم يشأ فرضها عن طريق الوحي ، كغيرها من الفرائض ، و إنما استدعى نبيه الكريم ، ليبين للناس جليلة القدر ، عظيمة المكانة ، و أنها مادة أساسية فى منهاج التربية الإسلامية ، فهى نظافة و نشاط و صحة و علم و أخلاق .

(ا)

١ - ما معنى الإسراء و المعراج ؟

٢ - ما موقف الناس عند قصة الإسراء و المعراج ؟

٣ - ما حكمة الإسراء و المعراج ؟

٤ - اذكر بعض الحوادث التي تفسر لنا أن الروح لها من القدرة على الجسم ما تستطيع به أن تؤثر عليه و تحتجزه ؟

٥ - ما حكمة الصلاة فى الإسلام ؟ (٥ درجات)

ب)

أعرب الكلمات التي تحتها خط (٥ درجات)

ج)

اضبط الفقرة الثالثة من القطعة بالشكل ضبطا تاما من (فالذي حدث ...إلى ... ما حكمة الإسراء و.المعراج)

(٥ درجات)

د) ترجم الفقرة الرابعة إلى اللغة الملايوية (٥ درجات)

ه) ترجم ما يأتي إلى اللغة العربية

" ايمان اداله سوات تناك يڠ بوله ممفرتهانكن ديري دري سكل روڤا كرنداهن دان كبوروقنن دان اي مندوروڠ كفد سكل روڤا كموليأن"

(٥ درجات)

المجموعة الثانية : المقال

(١٥ درجة)

اكتب مقالا فى أحد الموضوعات الآتية بحيث لا يقل عدد كلماته عن مائتي (٢٠٠) كلمة .

١ - دور الشباب فى بناء المستقبل المشرق للوطن

٢ - اكتب رسالة إلى زميل لك فى خارج البلاد تصف له سكان بلدك أيام رمضان .

٣ - " الإنسان مَدَنِيّ بالطبع " اكتب ما فهمت من هذا الكلام .

٤ - المواصفات للداعي إلى سبيل الله بالحكمة و الموعظة الحسنة.

٥ - صف رحلة قمت بها فى فترة العطلة الماضية .

القسم الثالث : الأسئلة الموضوعية

أجب عن الأسئلة الآتية فى بطاقة الحاسب الآلي المعدة لها :

(٢٠ درجة)

١ - (و من أضل ممن يدعو من دون الله من لا يستجيب له إلى يوم القيامة) الأية .

تستعمل من التي تحتها خط فى تلك الآية ل.....

A - العاقل إذا نزل غير العاقل منزلة العاقل

B - غير العاقل إذا نزل غير العاقل منزلة العاقل

C - العاقل إذا نزل العاقل منزلة غير العاقل

D - غير العاقل و قصد تغليب غير العاقل لكثرته فى اقترانهما

E - العاقل و قصد تغليب العاقل لأهميته فى اقترانهما

٢ - يبنى الفعل الماضى على الفتح إذا اتصلت به :

I - ألف الاثنين

II - واو الجماعة

III - تاء التأنيث الساكنة

IV - ياء المخاطبة

V - نون النسوة

A - III&II,I

B - V & IV,I

C - IV & III,I

D - III &I

E - IV & I

٣ - الكلمات التى تحتها خط فيما يأتي تعرب بالإعراب المحلي إلا :

A - رأيت الطالب يضحك

B - رأيت طالبا يضحك

C - قال أبي : القراءة مفتاح النجاح

D - الإسلام يدعو إلى العدل و المساواة

E - إن زيدا أخوه كريم

٤ - كل ما يأتي مواضع الضمير المستتر جوازا إلا :

A - فعل الأمر للواحد المفرد

B - الفعل الماضي للغائب المفرد

C - الفعل المضارع للغائب المفرد

D - الفعل الماضي للغائبة المفردة

E - الفعل المضارع للغائبة المفردة

٥ - تسمى " مِنْ " التي تحتها خط فى قولنا :

(هل من مخلصٍ يفعل ذلك ؟) ب.....:

A - الجر

B - حرف الجر الشبيه

C - حرف الجر الشبيه بالزائد

D - حرف الجر الزائد

E - حرف الجر الأصلي

٦ - المانع من ظهور الضمة ، و الفتحة على "يدعو " و "الفتى " و " الجاني " فى قولنا :

يدعو الفتى الجاني

٩ .

A - التعذر/ الثقل / التعذر

B - الثقل / التعذر / الثقل

C - التعذر/ التعذر / الثقل

D - الثقل / الثقل / التعذر

E - التعذر/ الثقل / الثقل

٧ - إعراب كلمة " محاميٍّ " في قولنا (مَرَرْتُ بِمُحَامِيٍّ) :

A - مجرور بالفتحة الظاهرة

B - مجرور بالفتحة المقدرة للثقل

C - مجرور بالكسرة المقدرة لسكون الإدغام

D - مجرور بالكسرة المقدرة لحركة الياء المناسبة

E - مجرور بالكسرة المقدرة للثقل

٨ - كل ما يأتي من النكرات التي لا تقبل (ال) و لا تقع موقع ما يقبل (ال) إلا :

A - حضر الطالب المحاضرة

B - أقابله ماشياً

C - لا طالب موجود

D - رأيت أحد عشر طالبا

E - قام الطالب إكراما للأستاذ

٩ - زمن الفعل المضارع الذي تحته خط في قوله تعالى :

(الوالدات يرضعن أولادهن حولين كاملين) الآية

A - الماضي

B - الحال

C - المستقبل

D - الماضي و الحال

E - الحال و المستقبل

١٠ - عين أقسام المعرفة مما يأتي :

I - اسم الموصول

II - اسم الإشارة

III - المنادى النكرة المقصودة

IV - الضمير

V - المضاف

A - V,IV,III,II,I

B - IV,III,II,I

C - V,III,II,I

D - V,IV,III,I

E - V,IV,II,I

١١ - عين الصحيح فيما يأتي :

A - الناقص هو ما كان وسطه حرف العلة

B - المهموز هو ما كان أوله و ثالثه حرف العلة

C - الأجوف هو ما كان وسطه حرف العلة

D - المثال هو ما كان آخره حرف العلة

E - المضعّف هو ما كان أوله حرف العلة

١٢ - عين الفعل المتعدي فيما يأتي :

A - صعب

B - قفز

C - أدْخَل

D - اتّصَل

E - انقطع

١٣ - الفعل " تقاتلوا " زِيْدَ فيه حرفان ، و هما :

A - التاء و الواو

B - التاء و الألف و الواو

C - التاء و التضعيف

D - التاء و الألف

E - الألف و الواو

١٤ - سَكَنْتَ فى الدور الرابع من هذه .....

A - المعمورة

B - العامرة

C - العمورة

D - العمارة

E - المعمارية

١٥ - الفعل المتعدي على وزن " افتعل " يفيد :

A - المشاركة

B - التعدية

C - الدخول فى شيء

D - الصيرورة

E - المطاوعة

١٦ - رتب الكلمات الآتية لتكون جملة مفيدة :

(دون / نزلت / التي / بالجهاد / موجهة / إلى / النساء / تأمر / القرآن / آيات / الرجال)

A - آيات القرآن تأمر بالجهاد موجهة التي نزلت إلى الرجال دون النساء

B - تأمر آيات القرآن نزلت التي موجهة إلى الرجال دون النساء بالجهاد

C - نزلت آيات القرآن التي تأمر بالجهاد موجهة دون النساء إلى الرجال

D - نزلت آيات القرآن التي تأمر بالجهاد موجهة إلى الرجال دون النساء

E - دون نزلت التي بالجهاد موجهة إلى النساء تأمر القرآن آيات الرجال

١٧ - (لم أتمكّنْ من .... تذكرة لحفل اليوم)

املأ الفراغ بالكلمتين المناسبتين :-

A - أحصل ، ل

B - حصول ، من

C - الحصول ، على

D - الحصيلة ، على

E - الحصول ، ل

١٨ - ترجم العبارة الآتية إلى اللغة الماليزية :

(جاء الطالب متأخراً)

A - فلاجر ايت داتغ ليوات

B - ليوتله داتغ فلاجر ايت

C - داتغله ليوات فلاجر ايت

D - ايت فلاجر يغ ليوات داتغ

E - تله داتغ اوله فلاجر ايت دالم كأدأن ليوات

١٩ - كيف تُعَبِّرُ هذه العبارة باللغة العربية :

(ساي اد سبواه كريتا بارو)

A - عندي سيارة التي جديدة

B - عندي السيارة الجديدة

C - عندي السيارة التي الجديدة

D - عندي سيارة جديدة

E - أنا موجود سيارة جديدة

٢٠ - كلمة من الكلمات الآتية لا علاقة بالأخرى :

A - أَبُوَيَ

B - أَبِيْ

C - أباهُ

D - أَبُوَّة

E - إباء

(مع تمنياتنا بالنجاح و التوفيق)

A.2. Test and examination papers

A.2.1 First Draft Placement test

بسم الله الرحمن الرحيم

اختبار تحديد المستوى للغة العربية : القراءة والمطالعة

الملاحظات :

١. أمامك دفتر للاسئلة وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة .

٣. يحتوي هذا الاختبار على ثلاثة أقسام : لكل قسم تعليمات خاصة للإجابة عن الأسئلة ، اقرأ التعليمات قبل أن تبدأ بالإجابة .

٤. الزمن المخصص للإجابة عن الأسئلة في هذا الاختبار خمسون دقيقة . والزمن المقترح لكل قسم مكتوب في بداية كل قسم .

٥. لكل سؤال درجة واحدة ولا تحاسَب على إجابة خاطئة .

٦. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة الأسئلة .

توقف الآن

لاتفتح ورقة الأسئلة حتى يسمح لك بذلك

القسم الأول (١٠ دقائق)

اقرأ كل فقرة ثم أجب عن الأسئلة التي تليها . اكتب إجابتك في ورقة الإجابة .

سئل أحد الخطباء عن الزمن الذي يحتاج إليه لإعداد خطبة يلقيها لمدة عشر دقائق فأجاب : أسبوعين . فقال السائل : فكيف إذا كانت الخطبة التي ستلقيها تحتاج إلى ساعة من الزمن ، فما الوقت الذي تحتاج إليه لإعدادها؟ فرد الخطيب : أسبوع واحد . فسأل السائل السؤال الثالث : فكيف إذا كانت الخطبة تحتاج إلى ساعتين في إلقائها ، فما الوقت الذي تحتاج إليه لإعدادها؟ فأجاب الخطيب : مثل هذه الخطبة لا تحتاج إلى إعداد ، وأنا على استعداد لإلقائها الآن !!!

١. كم من الوقت يحتاج إليه الخطيب لإعداد خطبة تستغرق ساعتين ؟

ا. أكثر من أسبوعين

بـ . أقل من أسبوع

جـ . لا يزيد عن ساعتين

د . لا يحتاج إلى وقت

٢. من هذا الحوار نفهم أن الخطيب يحتاج إلى وقت قصير لإعداد

ا. خطبة قصيرة

بـ . خطبة طويلة

جـ . خطبة قصيرة وطويلة

د . خطبة قصيرة وإلقائها

٣. نفهم من هذا الحوار كذلك أن الخطيب يرتجل ويجيد في

ا. إعداد خطبة

بـ . إلقاء خطبة طويلة

جـ . إجابة أسئلة

د . إلقاء خطبة قصيرة

اعلم أيها القارئ أن للمدخن " ثلاث فوائد " !! الأولى منها أن شعر المدخن لا يشيب ؛ والثانية الكلب يخاف من المدخن ؛ والثالثة اللص لا يدخل بيت المدخن . وتفسير ذلك أن شعر المدخن لا يشيب لأن المدخن عادة يموت مبكرا بسبب التدخين قبل أن يشيب شعره ! أما الكلب فإنه يخاف من المدخن لأن المدخن يتكئ على العصا عند المشي لمرض أصابه بسبب التدخين فيظن الكلب أن المدخن يريد أن يضربه ! وأما الفائدة الثالثة فإن اللص لا يدخل بيت المدخن لأن المدخن عادة أصابه السعال فلا يستطيع أن ينام فيظن اللص أنه يسهر في الليل !

٤. ماذا يقصد الكاتب من كلمة " فوائد " (الشطر الأول) في القطعة السابقة ؟

ا. منافع

بـ. خسائر

جـ. مفاسد

د. مغانم

٥. ينصحنا الكاتب في هذه القطعة بـ

ا. الإبتعاد عن التدخين لما له من خواطر

بـ. الإسراع إلى التدخين لما له من فوائد

جـ. الهروب من خطورة التشيب والكلب واللص

د. الابتعاد عن اللص إذا كان مدخنا

٦. الفكرة الاساسية في هذه القطعة هي أن التدخين

ا. ينفع المدخن بفوائد ثلاثة

بـ. يجعل عمر المدخن طويلا

جـ. يجعل المدخن لا ينام أبدا في الليل

د. يضر صحة المدخن

قال صلى الله عليه وسلم : (ما من مولود إلا يولد على الفطرة فأبواه يهودانه أو ينصرانه أو يمجسانه) . فالطفل أمانة عند والديه فإن عوداه الخير نشأ عليه وسعد في الدنيا والآخرة وإن أهملاه وتركاه لقرناء السوء شقي وهلك . فاعتدال الأطفال أو انحرافهم إنما يرجع إلى التنشئة والتربية .

٧. هذه الفقرة تبين لنا أهمية

ا. تربية الأولاد وتنشئتهم

بـ. قرناء الأولاد واعتقاداتهم

جـ. سعادة الأولاد في الدنيا والآخرة

د. اعتدال الأولاد وانحرافهم

٨. تبين الفقرة أن مستقبل الأطفال يعتمد على

ا. التربية التي اختارها لهم آباؤهم

بـ. الأعمال التي يقوم بها آباؤهم

جـ. القرناء الذين يعيشون حولهم

د. الأمانة التي تقع على آبائهم

فعلى الآباء أن يستمدوا من الإسلام مناهج التربية الصحيحة وذلك بأن يتخذوا من سيرة رسول الله صلى الله عليه وسلم القدوة الحسنة كما بينه سبحانه وتعالى في قوله { لقد كان لكم في رسول الله أسوة حسنة ...الآية } . وبذلك يكون الآباء أنفسهم قدوة حسنة لأبنائهم في القول والعمل .. وعليهم أن يربوهم مما رزقهم الله من مال حلال وأن يعلموهم من علوم الدين والدنيا ما يبتغون به طاعة الله ورسوله ويحقق لهم النفع والكسب في حياتهم . إذا اتبع الآباء هذا الأسلوب في تنشئة الأبناء يكونون قد جمعوا بين خيري الدنيا والآخرة وبنوا مجتمعهم على ركائز من الأخلاق والعلم والمال ...

٩. من عناصر التربية التي اقترحها الكاتب في هذه الفقرة هي

ا. اتخاذ سيرة الرسول صلى الله عليه وسلم كقدوة

II. أن يصبح الآباء نموذجا حسنا لأبنائهم

III. إطعام الآباء أولادهم طعاما حلالا طيبا

١٧. تعليم الآباء أبناءهم علوم الدين فقط

ا. ١

بـ. ١، II

جـ. II، III، IV

د. ١، II، III

١٠. يرى الكاتب أن الآباء الذين يتخذون الطريقة الصحيحة في تربية الأولاد قد أضمنوا لأولادهم

ا. الغناء والتقدم في العلوم الدنيوية

بـ. الإبتعاد عن مطالب الدنيا

جـ. القدرة على تفريق الدنيا من الآخرة

د. السعادة والنجاح في الدنيا والآخرة

==========

القسم الثاني (٢٠ دقيقة)

اقرأ كل فقرة ثم ضع علامة (✓) أمام عبارة صحيحة وعلامة (X) أمام عبارة خاطئة في ورقة الإجابة.

من المعلوم أن الصلاة إذا أديت كلها في الوقت المخصص لها فهي أداء، وإن فعلت مرة ثانية في الوقت لخلل غير الفساد فهي إعادة، وإن فعلت بعد الوقت فهي قضاء، والقضاء: فعل الواجب بعد وقته (من كتاب الفقه الإسلامي وأدلته جـ ١: ص ٥٦٦).

١١. إذا صلينا نفس الصلاة مرتين في الوقت المخصص لها والصلاة الأولى فاسدة أو باطلة فتسمى الصلاة الثانية إعادة.

١٢. نفهم من القطعة السابقة أن صلاة القضاء مخصصة للصلوات المفروضة فقط.

توجه الخليفة هارون الرشيد إلى المدينة المنورة وأراد أن يستمع إلى حديث من العالم الفقيه مالك بن أنس. فأرسل رسولا إليه يطلب منه الحضور إليه. فقال الإمام مالك للرسول: قل لأمير المؤمنين، إن طالب العلم يذهب إلى العلم، فأما العلم فلا يسعى إلى أحد. واقتنع الخليفة وزار العالم الفقيه في داره، لكنه أمر بإخلاء المكان من الناس. فرفض مالك وأصر أن يبقى الناس وقال: إذا منعنا العلم عن عامة الناس، فلا خير فيه للخاصة. ووافق الرشيد مرة أخرى على رغبة مالك، وسمح للناس بسماع الحديث معه.

١٣. هذه القصة تدلنا على أن العلم يؤتى ولا يأتي إلى طالبه

١٤. أراد الخليفة أن يكون مجلس العلم له وحده دون غيره لارتفاع مكانته من الناس

١٥. من هذه القصة نفهم كذلك أن الإمام مالك لا يحترم الخليفة عندما يرفض طلبه

كانت الأم لا تقرأ ولا تكتب. هربت في طفولتها من المدرسة لأنها كرهت مبادئ القراءة والكتابة. وبقيت معها هذه الكراهية حتى تزوجت. ولم تكن تشعر، بسبب هذا الجهل، بأي نقص في حياتها. فإذا أرادت أن تسمع الأخبار فتحت المذياع، وإذا أرادت محاسبة أحد استعانت بزوجها.

ثم أنجبت هذه الزوجة غير المتعلمة طفلة. وكبرت الطفلة ودخلت المدرسة. وفي أحد الأيام أحضرت الطفلة كراستها إلى أمها وطلبت منها أن تساعدها في تهجية إحدى الكلمات. وخجلت الأم أن تعترف لابنتها بأنها لا تقرأ ولا تكتب. كانت تحب ابنتها حبا لم تستطع فيه أن تواجهها بالحقيقة المرة. فذهبت على الفور إلى مدرسة ليلية وبدأت تتعلم القراءة والكتابة. كانت تسهر الليل لتلحق بدروس طفلتها.

وبدأت الأم تساعد ابنتها في مراجعة دروسها عاما بعد عام. واستمرت تتعلم دون علم ابنتها اثنتي عشرة سنة كاملة. وعندما جاء موعد امتحانات شهادة الدراسة الثانوية فوجئت الابنة بأن جارتها في امتحان الشهادة هي أمها، ودهشت الابنة فقد كانت تتصور أن أمها حصلت على هذه الشادة منذ اثني عشر عاما.

١٦. كانت الأم غبية في الدراسة أيام طفولتها وأصبحت ذكية بعد أن أنجبت بنتها.

١٧. التحقت الأم بمدرسة ليلية لأن زوجها تريدها أن تعلم بنتها في البيت .

١٨. درست الأم في مدرسة ليلية مقررات تختلف عن المقررات التي درستها بنتها في المدرسة .

١٩. دهشت الابنة عندما علمت أن أمها لم تحصل على شهادة الدراسة الثانوية بعدُ .

٢٠. من العبر في هذه القصة هي أن الإنسان يستطيع أن يفعل شيئا إذا كانت لديه رغبة قوية .

لست أريد من المحافظة على الوقت أن يملأ الوقت كله بالعمل ، وأن تكون الحياة كلها عملا لا راحة فيها وأن تكون عابسة لا ضحك فيها . فقد كان هذا للأسف هو المثل الأعلى في القرون الوسطى ، وكان خير الناس في ذلك الوقت من جد ولم يلعب ، وعبس ولم يضحك واستحضر الموت في كل لحظة . فلم تدخل السعادة قلبه ، ورآه الناس حزينا دائما كأنه راجع من زيارة الجنازة . وكان من خير ما دعا إليه العلماء في هذا العصر الحديث السرور والضحك واللعب في معقول من الوقت ، فذلك ينفع الناس أكثر من الجد الدائم .

أريد ألا تكون أوقات الفراغ طاغية على أوقات العمل ، وألا تكون أوقات الفراغ هي صميم الحياة ، وأوقات العمل على هامشها ، بل أريد أكثر من ذلك أن تكون أوقات الفراغ خاضعة لحكم العقل كأوقات العمل ، فإن كنا في العمل نعمل للغاية والهدف ، فيجب أن نصرف أوقات الفراغ لغاية كذلك ، إما لفائدة صحية كالألعاب الرياضية ، وإما للذة نفسية كالقراءات العلمية والأدبية .

أما أن تكون الغاية هي قتل الوقت ، فليست غاية مشروعة لأن الوقت هو الحياة فقتل الوقت قتل الحياة . فالذين يصرفون أوقاتهم الطويلة في لعبة شطرنج لا يعملون لغاية يرضاها العقل ، وكذلك الذين يتجولون بين المقاهي والأندية والطرقات لا يطلبون إلا قتل الوقت فإنهم أعداء للوقت .

٢١. يريد منا الكاتب في هذه القطعة أن نملأ أوقاتنا دائما بالعمل الجادّ

٢٢. يرى الكاتب أن الناس في القرون الوسطى يقسمون أوقاتهم لحياتهم تقسيما خاطئا

٢٣. يرى الكاتب أننا لا نحتاج إلى وضع الهدف للنشاطات في أوقات الفراغ وذلك لأننا قد وضعناه في أعمالنا

٢٤. لا يوافق الكاتب الذين يجعلون أوقات فراغهم أهمّ من أوقات أعمالهم

٢٥. تعتبر لعبة شطرنج والتجول وغيرهما من الأعمال المشروعة شريطة أن تكون هذه الأعمال غرضها قتل الوقت

عن أبي هريرة رضي الله عنه قال ، (سأل رجل رسول الله صلى الله عليه وسلم فقال ، يا رسول الله إنا نركب البحر ونحمل معنا القليل من الماء فإن توضأنا به عطشنا أفنتوضأ بماء البحر فقال رسول الله صلى الله عليه ، هو الطهور ماؤه الحل ميتته) رواه الخمسة .

الحديث أخرجه ابن خزيمة وابن حبان في صحيحيهما وابن الجارود في المنتقى والحاكم في المستدرك والدارقطني والبيهقي في سننهما وابن أبي شيبة . وحكى الترمذي عن البخاري تصحيحه وتعقُّبه ابن عبد البر بأنه لو كان صحيحا عنده لأخرجه في صحيحه . ثم حكم ابن عبد البر مع ذلك بصحته لتلقي العلماء له بالقبول فرده من حيث الاسناد وقبله من حيث المعنى . وصححه أيضا ابن المنذر وابن منده والبغوي وقال هذا الحديث صحيح متفق على صحته . وقال ابن الأثير في شرح المسند هذا حديث صحيح مشهور أخرجه الأئمة في كتبهم واحتجوا به ورجاله ثقات ... (مقتطف من كتاب نيل الأوطار للشوكاني بتصرف ، ص . ١٧)

٢٦. يدل الحديث السابق على أنه يجوز الوضوء بماء البحر وأن ميتة البحر حلال

٢٧. الحديث المذكور أعلاه لا يخرجه الدارقطني والبيهقي في سننهما

٢٨. ذكر الترمذي أن البخاري لا يخرج هذا الحديث لأنه لا يعتبره صحيحا

٢٩. الذي يرفض الحديث السابق من حيث الإسناد ويقبله من حيث المعنى هو ابن عبد البر

٣٠. احتجّ ابن الأثير بعدم صحة هذا الحديث في كتاب شرح المسند

==========

القسم الثالث (٢٠ دقيقة)

املأ الفراغات الآتية بكلمة مناسبة من عندك في ورقة الإجابة .

القطعة الأولى :

قال عز من قائل : { سبحان الذي أسرى بعبده ليلا من المسجد الحرام إلى المسجد الأقصى الذي باركنا حوله لنريه من آياتنا إنه هو السميع البصير .} الإسراء.

لقد أكرم الله نبيه في هذه الرحلة المباركة إلى السموات العلى وسدرة المنتهى حيث أراه من الآيات الكبرى ، وتجاوز السموات السبع حتى سدرة المنتهى . نعمْ لقد كانت رحلة ـــ٣١ـــ والمعراج دستورا راقيا للنجاة ، و ـــ٣٢ـــ لتخطي عقبات ومشكلات ، وإعانة ـــ٣٣ـــ بالنفس والنفيس في سبيل ـــ٣٤ـــ . لقد أرضاه ربه واجتباه و ـــ٣٥ـــ بالإسراء والمعراج . فإذا حال ـــ٣٦ـــ حوله ومرت الذكرى الكريمة ـــ٣٧ـــ عليهم أن يتدارسوا آثارهم ـــ٣٨ـــ بينهم ويستخلصوا العبر والعظات ـــ٣٩ـــ بما فيه خير الدنيا و ـــ٤٠ـــ .

وإذ تمر هذه الذكرى ـــ٤١ـــ اليوم بالمسلمين وهم في ـــ٤٢ـــ وتنافر وتخاصم وتناحر وجب ـــ٤٣ـــ أن يعملوا بكل جدية و ـــ٤٤ـــ على استعادة المسجد الأقصى ـــ٤٥ـــ إلى أخويه : المسجد الذي ـــ٤٦ـــ في الإسلام أولى القبلتين و ـــ٤٧ـــ الحرمين الشريفين ومسرى لرسولنا ـــ٤٨ـــ صلى الله عليه وسلم . لا بد من عودة هذا المسجد الذي تهفو إليه ـــ٤٩ـــ المسلمين من مشارق الأرض و ـــ٥٠ـــ . فلك الله يا أقصى ، ـــ٥١ـــ كان قدَّر لك أن ـــ٥٢ـــ أسيرا في يد الطغيان ـــ٥٣ـــ ، فإن نصر الله قريب ، و ـــ٥٤ـــ ذلك على الله ببعيد : ولئن ظلتْ حقبة من الزمن مكبلا بسلاسل الظلم فإن عين الله لا تنام .

القطعة الثانية :

لقد قصد الإسلام أن يكون الإنسان مثلا صالحا محمود الخصال ، شريف الشمائل ، كريم الأخلاق ، إن تكلم صدق ، وإن وعد ـــ٥٥ـــ بوعده ، وإن اؤتمن في الأمر ـــ٥٦ـــ الأمانة ولم يخن ، وإن رأى ـــ٥٧ـــ منكرا غيره بيده ، فإن لم ـــ٥٨ـــ فبلسانه ، فإن لم يستطع فبقلبه ، و ـــ٥٩ـــ تكلم خفض صوته ، وإن مشى ـــ٦٠ـــ يكن مختالا فخورا في مشيته ، و ـــ٦١ـــ رأى كبيرا وقره .

ومن الآداب و ـــ٦٢ـــ السلوك في الإسلام ما يلي :

ـــ٦٣ـــ المسلم أن يحسن الآداب في ـــ٦٤ـــ والمحادثة وأن يتلطف في التخاطب و ـــ٦٥ـــ الخشونة في الحديث ، قال تعالى : {وقولوا للناس حسنا} أي كلاما ـــ٦٦ـــ عند المحادثة والمخاطبة فيكون الحديث ـــ٦٧ـــ برفق ليس بالمرتفع ولا بالمنخفض ، و ـــ٦٨ـــ الأمور أوسطها ، ونهى الله عن ـــ٦٩ـــ في الكلام ، قال تعالى : { إن الذين يفترون على الله الكذب لا يفلحون } فالكاذب لا ينجح و ـــ٧٠ـــ يفلح في جميع أموره .

على ـــ٧١ـــ إن يؤدي التحية الحسنة ويفشي ـــ٧٢ـــ ، قال تعالى : { وإذا حييتم بتحية فحيوا بأحسن منها } . وعلى المسلم أن يوسع لجليسه ـــ٧٣ـــ أقبل عليه ، ويلتزم معه الأدب ـــ٧٤ـــ إذا كان أكبر منه سنا ، و ـــ٧٥ـــ إذا كان أبا أو أستاذا ـــ٧٦ـــ . وليس للقادم أن يقيم أحدا ـــ٧٧ـــ مجلس ليجلس مكانه ، قال صلى الله عليه وسلم : (لا يقيم الرجل الرجل من مجلسه ولكن تفسحوا وتوسعوا) .

وعلى المسلم ـــ٧٨ـــ يأكل حتى يجوع ، لأن الأكل ـــ٧٩ـــ الشبع مضرة أكيدة ، والإسلام يراعي ـــ٨٠ـــ الجسد وسلامته . وألا يأكل المسلم ـــ٨١ـــ يشرب إلا ما أحله الله . قال تعالى : {كلوا من طيبات ما رزقناكم} ـــ٨٢ـــ أراد المسلم أن يأكل فعليه

ـــ٨٢ـــ يطف يديه وفمه ، ثم يسمي بـ ـــ٨٤ـــ الله ، ويبدأ في الأكل بيمينه و ـــ٨٥ـــ . والأفضل للمسلم أن يأكل باليد ـــ٨٦ـــ ومما يليه . وبعد الانتهاء من ـــ٨٧ـــ يحمد المسلم ربه ويشمر على ـــ٨٨ـــ الأكل أو الشرب اقتداء برسول ـــ٨٩ـــ صلى الله عليه وسلم . ثم يغسل يديه وفمه . والإسلام ـــ٩٠ـــ دائما إلى ما فيه الخير ، و ـــ٩١ـــ للتنظيم ، فقد خصص اليد اليمنى ـــ٩٢ـــ في الأشياء الطيبة الكريمة ، مثل ـــ٩٣ـــ والشرب والمصافحة وحمل المصحف الشريف و ـــ٩٤ـــ العلم ، واليد اليسرى لغير ذلك ، كـ ـــ٩٥ـــ وتنظيف الأذن وحمل النعلين . فالشيء ـــ٩٦ـــ يجب أن يستعمل فيه اليد ـــ٩٧ـــ . وكذلك الرِّجْلان فالرجل اليمنى تستعمل لـ ـــ٩٨ـــ في المساجد وعند لبس ـــ٩٩ـــ ، قال رسول الله صلى الله عليه وسلم :(إذا انتعل أحدكم فليبدأ باليمين وإذا نزع فليبدأ بالشمال).

مع أحر تمنياتي لكم بالتوفيق والنجاح

بسم الله الرحمن الرحيم

اختبار تحديد المستوى للغة العربية : القواعد العربية

الملاحظات :

١. أمامك دفتر للاسئلة وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة .

٣. يحتوي هذا الاختبار على قسمين : لكل قسم تعليمات خاصة للإجابة عن الأسئلة ، اقرأ التعليمات قبل أن تبدأ بالإجابة .

٤. الزمن المخصص للإجابة عن الأسئلة في هذا الاختبار أربعون (٤٠) دقيقة . والزمن المقترح لكل قسم مكتوب في بداية كل قسم .

٥. لكل سؤال درجة واحدة ولا تحاسَب على إجابة خاطئة .

٦. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة الأسئلة .

توقف الآن

لاتفتح ورقة الأسئلة حتى يسمح لك بذلك

القسم الأول (٣٠ دقيقة)

يتكون كل سؤال في هذا القسم من جملة تنقصها كلمة أو عبارة. وبعد كل جملة توجد أربع كلمات أو عبارات . اختر الكلمة أو العبارة التي في رأيك تكون إجابة صحيحة . ثم ضع دائرة حول الحرف الذي يدل على الكلمة أو العبارة التي اخترتها في المكان المخصص في ورقة الإجابة .

١. رأيتُ الطالبات ..... في دراستهن .
ا. يجتهدون
بـ. تجتهدن
جـ. تجتهدين
د. يجتهدن

٢. يسكن في بيت أخي ..... كريم
ا. رجلان
بـ. رجلا
جـ. رجل
د. رجال

٣. ..... أستاذان بأكاديمية الدراسة الإسلامية وهما عضوان في مجلس الجامعة .
ا. فاطمة وزينب
بـ. إبراهيم وزينب
جـ. خديجة
د. علي ومحمد ويحي

٤. الطلبة والطالبات ..... احتفال العيد الوطني في العاصمة .
ا. يحضرون
بـ. يحضرن
جـ. يحضران
د. تحضران

٥. دخل النصراني في الإسلام فيصبح هو الآن ..... في الدين .
ا. أخينا
بـ. أخانا
جـ. أخونا
د. إخواننا

٦. إن الله يمكّن للمسلمين دينهم ..... ارتضى لهم .
ا. الذين
بـ. التي
جـ. اللذين
د. الذي

٧. تقرأ كلمة " فاطمة " في الجملة : إنّ فاطمة طالبة مجتهدة ..بـ :
ا. الكسرة
بـ. الضمة
جـ. الفتحة
د. الفتحتين

٨. ظن سكان القرية أنّ ..... محمد قد غرق في الماء .
ا. أخو
بـ. أخ
جـ. أخي
د. أخا

٩. قلت لنفسي عندما توفي زميلي محمد : يا ليت ....
ا. محمدا موجودا
بـ. محمدا موجود
جـ. محمدين موجودين
د. محمد موجود

١٠. اسم كان في قوله تعالى : { لقد كان لكم في رسول الله أسوة حسنة ... الآية } هو :
ا. أسوة
بـ. رسول الله
جـ. الله
د. حسنَة

١١. إن الزوج ..... بما يحكم عليه القاضي في مشكلته الزوجية .
ا. راضيا
بـ. راض
جـ. راضيين
د. راضيان

١٢. أصبح .... ربهم عند الفجر .
ا. المسلمون داعون
بـ. المسلمين داعون
جـ. المسلمون داعين
د. المسلمين داعين

١٣. مثنى كلمة " منتدى " هو
ا. منتدوان
بـ. منتدان
جـ. منتديان
د. منتدءان

١٤. جذور كلمة " انحاز " هو
ا. حيز
بـ. حوز
جـ. نحز
د. انحز

١٥. فعل الأمر المفرد المؤنث لفعل " استعاذ " هو :
ا. استعيذي
بـ. استعذي
جـ. تعذي
د. استعوذي

١٦. الأمهات ..... الطعام في المطبخ .
ا. أعدت
بـ. أعد
جـ. أعدوا
د. أعددن

١٧. إن المسلمات .....ليالي رمضان إيمانا واحتسابا .
ا. قامت
بـ. قمن
جـ. قاموا
د. قامتا

١٨. فعل الأمر المفرد المذكر لـ " اتقى " هو
ا. اتقي
بـ . أق
جـ . وقي
د. اتق

١٩. بدأت عملي ولم ..... منه بعد .
ا. أنتهيا
بـ . أنتهيَ
جـ . أنته
د. أنتهي

٢٠. أنتما لا ..... إلا بالخير .
ا. تدعوا
بـ . تدعوان
جـ . تدعو
د. تدعون

٢١. إننا نحترمك لأنك تأمر المسلمين بالمعروف و ..... عن المنكر .
ا. تنهونه
بـ . تنهانا
جـ . تنهاهم
د. تنهوننا

٢٢. أشفقت على المرأتين ..... فقدتا أبناءهما .
ا. اللتان
بـ . اللواتي
جـ . التي
د. اللتين

٢٣. المبتدأ في الجملة : " في مساجد رجال كثيرون يذكرون الله " هو
ا. مساجد
بـ . رجال
جـ . الله
د. كثيرون

٢٤. قالت الطالبة : أنا ..... إلى الجامعة غدا .
ا. حاضر
بـ . حاضرات
جـ . حاضرا
د. حاضرة

٢٥. فرض الله الزكاة على الأغنياء . فكان ..... أموال يدفعون زكاتهم ابتغاء مرضاة الله .
ا. ذوُو
بـ . ذوًا
جـ . ذو
د. ذوِي

٢٦. أيها المؤمنون ..... يدا واحدة في المنشط والمكره .
ا. كونوا
بـ . تكونون
جـ . كوني
د. كن

٢٧. سمعت أن ..... علي قد وصلا من السفر .
ا. أخوان
بـ . أخوين
جـ . أخوي
د. أخوا

٢٨. إن ..... بايعن الرسول وآمنَ وعملن الصالحات سيدخلن الجنة .
ا. اللاتي
بـ . اللتين
جـ . التي
د. اللتان

٢٩. بعدما شرح المعلم ، لا تزال .... في رأيي .
ا. المسألتين غامضتين
بـ . المسألتين غامضتان
جـ . المسألتان غامضتان
د. المسألتان غامضتين

٢٦. يعرف المنافق أن الناس لا ..... ومع ذلك ..... في نفاقه .
ا. يحترمونه ..... يستمر
بـ. يحترمونه ..... يستمر
جـ. يحترمه ..... يستمرون
د. يحترمه ..... يستمر

٢٧. ..... المسلمون بدين الله الحنيف ولا يتفرقون .
ا. يتمسكون
بـ. تتمسكون
جـ. تتمسك
د. يتمسك

٢٨. استفدت من كتب أستاذي ..... استعرت منه .
ا. اللتين
بـ. الذي
جـ. التي
د. الذين

٢٩. للقضاة العادلين ..... كريم عند الله
ا. مقام
بـ. مقامون
جـ. المقام
د. مقامين

٤٠. ما إعراب كلمة " لآيات " في قوله تعالى : { إن في خلق السموات والأرض واختلاف الليل والنهار لآيات لأولي الألباب }
ا. خبر إن
بـ. اسم إن
جـ. الحال
د. التمييز

٤١. عاد إلى البلاد الولدان ..... سافرا إلى مكة .
ا. اللذين
بـ. الذين
جـ. اللذان
د. الذي

٢٠. مثنى كلمة " عصى " هو
ا. عصان
بـ. عصوان
جـ. عصين
د. عصيان

٢١. مثنى كلمة " حمراء " هو
ا. حمراوان
بـ. حمراءان
جـ. حمران
د. حمرايان

٢٢. جذور كلمة " انصرف " هو
ا. نصر
بـ. نرف
جـ. انصر
د. صرف

٢٣. فعل الأمر المفرد المؤنث لـ " أدارَ " هو
ا. أري
بـ. دري
جـ. أديري
د. أدري

٢٤. ما إعراب كلمة " أخوك " في الجملة الآتية : " كان من بين الفائزين في المسابقة أخوك "
ا. فاعل
بـ. اسم كان
جـ. خبر كان
د. مفعول به

٢٥. إعراب كلمة " راع " في الجملة : " كلكم راع وكلكم مسؤول عن رعيته " هو
ا. مضاف إليه
بـ. صفة لكلكم
جـ. خبر لكلكم
د. مبتدأ مؤخر

٤٨. ..... في شهر رمضان وستة أيام من شوال .

ا. صام
بـ. صاموا
جـ. صمت
د. صمنا

٤٩. جذور كلمة " اصطبر " هو
ا. طبر
بـ. صبر
جـ. اصبر
د. صطبر

٥٠. قال الطالب لشيخه : إني ..... قومي ليلا ونهارا .
ا. دعا
بـ. دعا أنا
جـ. دعيت
د. دعوت

القسم الثاني (١٠ دقائق)

اقرأ العبارات الاتية . إذا كانت العبارة في رأيك صحيحة ضع علامة (✓) في ورقة الإجابة وإذا كانت العبارة في رأيك خاطئة ضع علامة (×) في ورقة الإجابة ثم اكتب الإجابة الصحيحة كما في المثالين الآتيين :

المثال الأول :

الجملة الآتية جملة فعلية : ذهب محمد إلى المسجد
العبارة " الجملة الآتية جملة فعلية " صحيحة ولذلك وضعت علامة (✓) في ورقة الإجابة .

المثال الثاني :

الجملة الآتية جملة فعلية : محمد يراجع درسه
العبارة " الجملة الآتية جملة فعلية " خاطئة ولذلك وضعت علامة (×) في ورقة الإجابة ثم كتبت العبارة الصحيحة وهي : الجملة الآتية جملة اسمية (الإجابة المختصرة كقولك : جملة اسمية تكون مقبولة كذلك) .

٤٢. تُقرأ كلمة " قوية " في الجملة : زينب قصيرة لكنها قوية بـ
ا. الفتحة الظاهرة
بـ. الضمة الظاهرة
جـ. الفتحة المقدرة
د. الكسرة الظاهرة

٤٣. أرى أن .... من أهم كتب المراجع للفقه الإسلامي .
ا. هذان الكتابان
بـ. هذين الكتابان
جـ. هذين الكتابين
د. هذان الكتابين

٤٤. إن ..... البخاري ومسلم من الكتب المعروفة في علم الحديث .
ا. صحيحان
بـ. صحيحين
جـ. صحيح
د. صحيحي

٤٥. اسم كان في قوله تعالى : { فما كان جوابَ قومه إلا أن قالوا اقتلوه أو حرقوه ... الآية } هو
ا. أن قالوا
بـ. قومه
جـ. جواب
د. اقتلوه

٤٦. اسم كان في قوله تعالى : { ما كان على النبي من حرج فيما فرض الله ... الآية } هو
ا. النبي
بـ. حرج
جـ. فرض
د. الله

٤٧. جذور كلمة " اشتاق " هو
ا. شيق
بـ. شتق
جـ. اشتق
د. اشتا

٥١. قال تعالى { ألم نجعل له عينين ولسانا وشفتين } .

إعراب كلمة (عينين) في الآية السابقة مفعول به منصوب بالياء .

٥٢. اسم الإشارة واسم الموصول والضمير من الأسماء المعرفة

٥٣. صيغة الأمر لفعل (اتقى) في حالة الجمع هي (اتقُوا) .

٥٤. كلمة (رجل) في الجملة : " في الدار رجل " هي الخبر

٥٥. تقرأ كلمتا (عبرة كثيرة) في الجملة : " كان لنا في قصص الأمم السابقة عبرة كثيرة " بالفتحة الظاهرة .

٥٦. إعراب كلمة (هناك) في الجملة : " هناك رجل يبحث عن أخيه " هو الخبر المقدم .

٥٧. تثنية كلمتي (الفتى) و (العصى) هي (الفتيان) و (العصوان) .

٥٨. جذور الكلمة لفعل (اشتق) هو (شتق)

٥٩. الجملة الآتية جملة فعلية : " أن يراجع محمد درسه خير له من أن يشاهد التلفزيون "

٦٠. تثنية كلمتي (الصحراء) و (الحمراء) هي (الصحران) و (الحمران) .

٦١. كلمة (ذهب) مثال للفعل الصحيح وكلمة (شدّ) مثال للفعل المضعف وكلمة (وعد) مثال للفعل المعتل .

٦٢. أحد تصريف فعل (صفى) هو (اصطفى) .

٦٣. جذور الكلمة لفعل (اتخذ) هو (خذ)

٦٤. صيغتا الأمر في حالة الإفراد للفعل الرباعي (ساوى) و (عادى) هما (ساوِ) و (عادِ)

٦٥. خبر ليت في قوله تعالى : { يليت لنا مثل ما أوتي قارون ... الآية } هو (مثل) .

مع أحر تمنياتي لكم بالتوفيق والنجاح

بسم الله الرحمن الرحيم

اختبار تحديد المستوى فى اللغه العربية : المقال

الملاحظات

١ أمامك دفتر للسؤال وورقة منفصلة للإجابة .

٢ اكتب البيانات المطلوبة في ورقة الإجابة .

٣ الزمن المخصص لكتابة المقال في هذا الاختبار ثلاثون دقيقة فقط .

٤ الدرجة الكاملة خمسون درجة .

٥ اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة السؤال .

توقف الآن

لاتفتح ورقة السؤال حتى يسمح لك بذلك

اكتب مقالة قصيرة تحت موضوع " **الالتحاق بالجامعة** " مستعينا بالأفكار الآتية

- حصولك على نتيجة امتحان الشهادة التوجيهية (STPM) أو نتيجة الامتحان الأخير في القسم التأهيلي (pra-akademi).

- تقديم الطلب للالتحاق بالجامعة (طلب الاستمارة ، إملاؤها ، إرسالها إلى جهة معينة).

- حصولك على القبول لمواصلة الدراسة في الجامعة (استعدادات للسفر كتب ، ملابس ، شهادات وغيرها) .

- قدومك إلى الجامعة (أول يوم في الجامعة ، الحياة في الجامعة) .

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : الإملاء

ستمع جيدا إلى التوجيهات الآتية

هذا اختبار الإملاء . ستستمع إلى قطعة تُقرأ ثلاث مرات . عندما تقرأ القطعة للمرة الأولى استمع فقط ولا تكتب شيئا على ورقة الإجابة . وعندما تقرأ القطعة للمرة الثانية حيث أن تكتبها في ورقة الإجابة . وستقرأ القطعة في هذه المرة بقراءة بطيئة في مقاطع يتسنى لك كتابتها . ثم ستقرأ القطعة للمرة الأخيرة للمراجعة وبإمكانك أن تكتب ما فقده في هذه المرة . وبعد ذلك تخصص دقيقتان للمراجعة الأخيرة .

الملاحظة

عندما تسمع الكلمات مثل فصل و وقف ونقطتان فلا تكتبها وإنما ضع علامتها فقط .

ستعد الآن لنبدأ بالقراءة الأولى . استمع جيدا ولا تكتب شيئا على ورقة الإجابة

قال رسول الله صلى الله عليه وسلم / { من رأى منكم منكرا / فليغيره بيده / إن لم يستطع فبلسانه / فإن لم يستطع فبقلبه / وذلك أضعف الإيمان }.

نفهم من هذال الحديث / أن النهي عن المنكر / واجبٌ / على كل مسلم ومسلمة /. هذا النهيُ / يقع في ثلاثِ مراحلَ / أعلاها / أن يمنع مسلمٌ منكرا بيده / أي بقدرته / كسر زجاجة الخمر / ومنع الظّالم / من أن يضرب / أو يؤذيَ المظلومَ / . فإذا لم يستطع المسلمُ / أن يفعل ذلك / لضعفه أو للخطورة الّتي ستقع عليه / ، انتقل الأمرُ إلى المرحلة الثانية / وهي / أن يمنع المنكرات بلسانه / أي / بوعظه وخطبته وكتاباته / وغيرها منَ النّشاطات اللّسانيةُ / . فإذا لم يقدر كذلكَ / بهذه الطريقة / ، انتقل الأمر إلى أدنى المراحل / وهي / المنع بالقلب / بحيث لا يرضى / عن المنكر الّذي يحدث أمامهُ / .

من هذا الحديث / نستنبط / أنه لا يجوز لمسلم / أن يرى منكرا دون أن يقوم بمنعه وباهتمام المسلمينَ بهذا الأمر النّبويّ / فقد ضمنوا لأنفسهم / السّعادةَ / في الحياة

القراءة للمرة الثانية للإملاء استمع جيدا واكتب ما تستمع إليه .

قال رسول الله صلى الله عليه وسلم / (١٠) (نقطتان) من رأى منكم منكرا /(١) فليغيره بيده / (٦) فإن لم يستطع فبلسانه (١٠) فإن لم يستطع فبقلبه / (١٠) وذلك أضعف الإيمانْ / . (١٠) (وقف)

نفهم من هذا الحديثْ / (٨) أن النهي عن المنكر / (٨) واجب على / (٣) كل مسلم ومسلمة / (٦) . (وقف) وهذا النهيُ / (٣) يقع في ثلاثِ مراحلَ / (١٠) (نقطتان) أعلاها / (٣) أن يمنع مسلم منكرا / (٨) بيده أي بقدرته / (٧) ككسر زجاجة الخمر / (٢) ومنع الظالم / (٤) من أن يضرب / (٦) أو يؤذي المظلومَ / (٦) . (وقف) فإذا لم يستطع المسلمُ / (١٠) أن يفعل ذلكَ / (٥) لضعفه أو للخطورة / (٦) التي ستقع عليه / (٦) ، (فصل) انتقلَ الأمر إلى المرحة الثانية / (١٣) وهي / (٣) أن يمنع المنكرات بلسانه / (١٣) أي بوعظه وخطبته / (٧) وكتاباته وغيرها / (٦) منَ النّشاطاتِ اللّسانية / (٧) . (وقف) فإذا لم يقدر كذلكَ / (٨) بهذه الطّريقة / (٤) انتقل الأمر إلى أدنى المراحل / (١٢) وهي / (٣) المنعُ بالقلب / (٧) بحيث لا يرضى عن المنكر / (٩) الذي يحدث أمامهْ / (٦) . (وقف) من هذا الحديثْ / (٦) نستنبط / (٣) أنه لا يجوز لمسلم / (٧) أن يرى منكرا / (٤) دون أن يقوم بمنعه / (٧) . (وقف) وباهتمام المسلمينَ بهذا الأمر النّبويّ / (١٥) فقد ضمنوا لأنفسهم / (٦) السعادةَ / (٣) في الحياة الدّنيوية والأخروية / (٩) . (وقف)

القراءة للمرة الثالثة : استمع جيدا وراجع ما كتبت من القطعة السابقة (راجع القطعة المقروءة للمرة الأولى).

(الملاحظة للمشرف : انتظر دقيقتين)

A.2.2 Second Draft Placement test

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : القراءة والمطالعة

الملاحظات :

١. أمامك دفتر للاسئلة وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة .

٣. يحتوي هذا الاختبار على ثلاثة أقسام : لكل قسم تعليمات خاصة للإجابة عن الأسئلة ، اقرأ التعليمات قبل أن تبدأ بالإجابة .

٤. الزمن المخصص للإجابة عن الأسئلة في هذا الاختبار خمسون دقيقة . والزمن المحدد لكل قسم مكتوب في بداية كل قسم .

٥. لكل سؤال درجة واحدة ولا تحاسَب على إجابة خاطئة .

٦. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة الأسئلة .

توقف الآن

لاتفتح ورقة الأسئلة حتى يسمح لك بذلك

القسم الأول (١٠ دقائق)

اقرأ كل فقرة ثم أجب عن الأسئلة التي تليها . ضع دائرة حول الحرف الذي يدل على الإجابة الصحيحة في رأيك في ورقة الإجابة .

سئل أحد الخطباء عن الزمن الذي يحتاج إليه لإعداد خطبة يلقيها لمدة عشر دقائق فأجاب : أسبوعين . فقال السائل : فكيف إذا كانت الخطبة التي ستلقيها تحتاج إلى ساعة من الزمن ، فما الوقت الذي تحتاج إليه لإعدادها؟ فرد الخطيب : أسبوع واحد . فسأل السائل السؤال الثالث : فكيف إذا كانت الخطبة تحتاج إلى ساعتين في إلقائها ، فما الوقت الذي تحتاج إليه لإعدادها؟ فأجاب الخطيب : مثل هذه الخطبة لا تحتاج إلى إعداد ، وأنا على استعداد لإلقائها الآن !!!

١ . كم من الوقت يحتاج إليه الخطيب لإعداد خطبة تستغرق ساعتين ؟

ا. أكثر من أسبوعين .

بـ . أقل من أسبوع .

جـ . لا يزيد عن ساعتين .

د . لا يحتاج إلى وقت .

٢ . من هذا الحوار نفهم أن الخطيب يحتاج إلى وقت قصير لإعداد

ا. خطبة قصيرة .

بـ . خطبة طويلة .

جـ . خطبة قصيرة وطويلة .

د . خطبة قصيرة وإلقائها .

٣ . نفهم من هذا الحوار كذلك أن الخطيب يرتجل ويجيد في

ا. إعداد خطبة .

بـ . إلقاء خطبة طويلة .

جـ . إجابة أسئلة .

د . إلقاء خطبة قصيرة .

اعلم أيها القارئ أن للمدخن " ثلاث فوائد " !! الأولى منها أن شعر المدخن لا يشيب ؛ والثانية الكلب يخاف من المدخن ؛ والثالثة اللص لا يدخل بيت المدخن . وتفسير ذلك أن شعر المدخن لا يشيب لأن المدخن عادة يموت مبكرا بسبب التدخين قبل أن يشيب شعره ! أما الكلب فإنه يخاف من المدخن لأن المدخن يتكئ على العصا عند المشي لمرض أصابه بسبب التدخين فيظن الكلب أن المدخن يريد أن يضربه ! وأما الفائدة الثالثة فإن اللص لا يدخل بيت المدخن لأن المدخن عادة أصابه السعال فلا يستطيع أن ينام فيظن اللص أنه يسهر في الليل !

٤ . ماذا يقصد الكاتب من كلمة " فوائد " (الشطر الأول) في القطعة السابقة ؟

ا. منافع .

بـ . خسائر .

جـ . مفاسد .

د . مغانم .

٥ . ينصحنا الكاتب في هذه القطعة بـ

ا. الإبتعاد عن التدخين لما له من مضار .

بـ . الإسراع إلى التدخين لما له من فوائد .

جـ . الهروب من خطورة التشيب والكلب واللص .

د . الابتعاد عن اللص إذا كان مدخنا .

٦ . الفكرة الأساسية في هذه القطعة هي أن التدخين

ا. ينفع المدخن بفوائد ثلاثة .

بـ . يجعل عمر المدخن طويلا .

جـ . يجعل المدخن لا ينام أبدا في الليل .

د . يضر صحة المدخن .

قال صلى الله عليه وسلم : (ما من مولود إلا يولد على الفطرة فأبواه يهودانه أو ينصرانه أو يمجسانه) . فالطفل أمانة عند والديه فإن عوداه الخير نشأ عليه وسعد في

الدنيا والآخرة وإن أهملاه وتركاه لقرناء السوء شقي وهلك . فاعتدال الأطفال أو انحرافهم إنما يرجع إلى التنشئة والتربية .

٧. هذه الفقرة تبين لنا أهمية :

ا. تربية الأولاد وتنشئتهم .

بـ . قرناء الأولاد واعتقاداتهم .

جـ . سعادة الأولاد في الدنيا والآخرة .

د . اعتدال الأولاد وانحرافهم .

٨. تبين الفقرة أن مستقبل الأطفال يعتمد على :

ا. التربية التي اختارها لهم آباؤهم .

بـ . الأعمال التي يقوم بها آباؤهم .

جـ . القرناء الذين يعيشون حولهم .

د . الأمانة التي تقع على آبائهم .

فعلى الآباء أن يستمدوا من الإسلام مناهج التربية الصحيحة وذلك بأن يتخذوا من سيرة رسول الله صلى الله عليه وسلم القدوة الحسنة كما بينه سبحانه وتعالى في قوله { لقد كان لكم في رسول الله أسوة حسنة ...الآية } . وبذلك يكون الآباء أنفسهم قدوة حسنة لأبنائهم في القول والعمل .. وعليهم أن يربوهم مما رزقهم الله من مال حلال وأن يعلموهم من علوم الدين والدنيا ما يبتغون به طاعة الله ورسوله ويحقق لهم النفع والكسب في حياتهم . إذا اتبع الآباء هذا الأسلوب في تنشئة الأبناء يكونون قد جمعوا بين خيري الدنيا والآخرة وبنوا مجتمعهم على ركائز من الأخلاق والعلم والمال ...

٩. من عناصر التربية التي اقترحها الكاتب في هذه الفقرة هي :

ا. اتخاذ سيرة الرسول صلى الله عليه وسلم كقدوة .

II. أن يصبح الآباء نموذجا حسنا لأبنائهم .

III. إطعام الآباء أولادهم طعاما حلالا طيبا .

IV. تعليم الآباء أبناءهم علوم الدين فقط .

ا. I

بـ . I ، II

جـ . II ، III ، IV

د . I ، II ، III

١٠. يرى الكاتب أن الآباء الذين يتخذون الطريقة الصحيحة في تربية الأولاد قد ضمنوا لأولادهم :

ا. الغناء والتقدم في العلوم الدنيوية .

بـ . الإبتعاد عن مطالب الدنيا .

جـ . القدرة على تفريق الدنيا من الآخرة .

د . السعادة والنجاح في الدنيا والآخرة .

==========

القسم الثاني (٢٠ دقيقة)

اقرأ كل فقرة ثم ضع علامة (✓) أمام عبارة صحيحة وعلامة (X) أمام عبارة خاطئة في ورقة الإجابة.

من المعلوم أن الصلاة إذا أديت كلها في الوقت المخصص لها فهي أداء ، وإن فعلت مرة ثانية في الوقت لخلل غير الفساد فهي إعادة ، وإن فعلت بعد الوقت فهي قضاء ، والقضاء : فعل الواجب بعد وقته (من كتاب الفقه الإسلامي وأدلته جـ ١ : ص ٥١٦) .

١١. إذا صلينا نفس الصلاة مرتين في الوقت المخصص لها والصلاة الأولى لم تكن باطلة فتسمى الصلاة الثانية إعادة .

١٢. نفهم من القطعة السابقة أن صلاة القضاء مخصصة للصلوات المفروضة فقط .

توجه الخليفة هارون الرشيد إلى المدينة المنورة وأراد أن يستمع إلى حديث من العالم الفقيه مالك بن أنس . فأرسل رسولا إليه يطلب منه الحضور إليه . فقال الإمام مالك للرسول : قل لأمير المؤمنين ، إن طالب العلم يذهب إلى العلم ، فأما العلم فلا يسعى إلى أحد . واقتنع الخليفة وزار العالم الفقيه في داره ، لكنه أمر بإخلاء المكان من الناس . فرفض مالك وأصر أن يبقى الناس وقال : إذا منعنا العلم عن عامة الناس ، فلا خير فيه للخاصة . ووافق الرشيد مرة أخرى على رغبة مالك ، وسمح للناس بسماع الحديث معه .

١٣. هذه القصة تدلنا على أن العلم يؤتى ولا يأتي إلى طالبه .

١٤. أراد الخليفة أن يكون مجلس العلم له وحده دون غيره لارتفاع مكانته من الناس .

١٥. من هذه القصة نفهم كذلك أن الإمام مالك لا يحترم الخليفة عندما يرفض طلبه .

كانت الأم لا تقرأ ولا تكتب . هربت في طفولتها من المدرسة لأنها كرهت مبادئ القراءة والكتابة . وبقيت معها هذه الكراهية حتى تزوجت . ولم تكن تشعر ، بسبب هذا الجهل ، بأي نقص في حياتها . فإذا أرادت أن تسمع الأخبار فتحت المذياع ، وإذا أرادت محاسبة أحد استعانت بزوجها .

ثم أنجبت هذه الزوجة غير المتعلمة طفلة . وكبرت الطفلة ودخلت المدرسة . وفي أحد الأيام أحضرت الطفلة كراستها إلى أمها وطلبت منها أن تساعدها في تهجية إحدى الكلمات . وخجلت الأم أن تعترف لابنتها بأنها لا تقرأ ولا تكتب . كانت تحب ابنتها حبا لم تستطع فيه أن تواجهها بالحقيقة المرة . فذهبت على الفور إلى مدرسة ليلية وبدأت تتعلم القراءة والكتابة . كانت تسهر الليل لتلحق بدروس طفلتها .

وبدأت الأم تساعد ابنتها في مراجعة دروسها عاما بعد عام . واستمرت تتعلم دون علم ابنتها اثنتي عشرة سنة كاملة . وعندما جاء موعد امتحانات شهادة الدراسة الثانوية فوجئت الابنة بأن جارتها في امتحان الشهادة هي أمها ، ودهشت الابنة فقد كانت تتصور أن أمها حصلت على هذه الشادة منذ اثني عشر عاما .

١٦. كانت الأم تحب الدراسة أيام طفولتها وأصبحت ذكية بعد أن تزوجت .

١٧. التحقت الأم بمدرسة ليلية لأن زوجها يريدها أن تعلّم بنتها في البيت .

١٨. درست الأم في مدرسة ليلية مقررات تختلف عن المقررات التي درستها بنتها في المدرسة .

١٩. دهشت الابنة عندما علمت أن أمها لم تحصل على شهادة الدراسة الثانوية بعدُ .

٢٠. من العبر في هذه القصة هي أن الإنسان يستطيع أن يفعل شيئا إذا كانت لديه رغبة قوية .

لست أريد من المحافظة على الوقت أن يملأ الوقت كله بالعمل ، وأن تكون الحياة كلها عملا لا راحة فيها وأن تكون عابسة لا ضحك فيها . فقد كان هذا للأسف هو المثل الأعلى في القرون الوسطى ، وكان خير الناس في ذلك الوقت من جد ولم يلعب ، وعبس ولم يضحك واستحضر الموت في كل لحظة . فلم تدخل السعادة قلبه ، ورآه الناس حزينا دائما كأنه راجع من زيارة الجنازة . وكان من خير ما دعا إليه العلماء في هذا العصر الحديث السرور والضحك واللعب في معقول من الوقت ، فذلك ينفع الناس أكثر من الجد الدائم .

أريد ألا تكون أوقات الفراغ طاغية على أوقات العمل ، وألا تكون أوقات الفراغ هي صميم الحياة ، وأوقات العمل على هامشها ، بل أريد أكثر من ذلك أن تكون أوقات الفراغ خاضعة لحكم العقل كأوقات العمل ، فإن كنا في العمل نعمل للغاية والهدف ، فيجب أن نصرف أوقات الفراغ لغاية كذلك ، إما لفائدة صحية كالألعاب الرياضية ، وإما للذة نفسية كالقراءات العلمية والأدبية .

أما أن تكون الغاية هي قتل الوقت ، فليست غاية مشروعة لأن الوقت هو الحياة فقتل الوقت قتل الحياة . فالذين يصرفون أوقاتهم الطويلة في لعبة شطرنج لا يعملون لغاية يرضاها العقل ، وكذلك الذين يتجولون بين المقاهي والأندية والطرقات لا يطلبون إلا قتل الوقت فإنهم أعداء للوقت .

٢١. يريد منا الكاتب في هذه القطعة أن نملأ أوقاتنا دائما بالعمل الجادّ .

٢٢. يرى الكاتب أن الناس في القرون الوسطى يقسمون أوقاتهم لحياتهم تقسيما خاطئا.

٢٣. يرى الكاتب أننا لا نحتاج إلى وضع الهدف للنشاطات في أوقات الفراغ وذلك لأننا قد وضعناه في أعمالنا .

٢٤. لا يوافق الكاتب الذين يجعلون أوقات فراغهم أهمّ من أوقات أعمالهم .

٢٥. تعتبر لعبة شطرنج والتجول وغيرهما من الأعمال المشروعة شريطة أن تكون هذه الأعمال غرضها قتل الوقت .

عن أبي هريرة رضي الله عنه قال ، (سأل رجل رسول الله صلى الله عليه وسلم فقال ، يا رسول الله إنا نركب البحر ونحمل معنا القليل من الماء فإن توضأنا به عطشنا أفنتوضأ بماء البحر فقال رسول الله صلى الله عليه وسلم ، هو الطهور ماؤه الحل ميتته) رواه الخمسة .

الحديث أخرجه ابن خزيمة وابن حبان في صحيحيهما وابن الجارود في المنتقى والحاكم في المستدرك والدارقطني والبيهقي في سننهما وابن أبي شيبة . وحكى الترمذي عن البخاري تصحيحه وتعقُّبه ابن عبد البر بأنه لو كان صحيحا عنده لأخرجه في صحيحه . ثم حكم ابن عبد البر مع ذلك بصحته لتلقي العلماء له بالقبول فرده من حيث الاسناد وقبله من حيث المعنى . وصححه أيضا ابن المنذر وابن منده والبغوي وقال هذا الحديث صحيح متفق على صحته . وقال ابن الأثير في شرح المسند هذا حديث صحيح مشهور أخرجه الأئمة في كتبهم واحتجوا به ورجاله ثقات ... (مقتطف من كتاب نيل الأوطار للشوكاني بتصرف ، ص ١٧٠)

٢٦. يدل الحديث السابق على أنه يجوز الوضوء بماء البحر وأن ميتة البحر حلال

٢٧. الحديث المذكور أعلاه ما أخرجه الدارقطني والبيهقي في سننهما .

٢٨. ذكر الترمذي أن البخاري ما أخرج هذا الحديث لأنه لا يعتبره صحيحا .

٢٩. الذي يرفض الحديث السابق من حيث الإسناد ويقبله من حيث المعنى هو ابن عبد البر .

٣٠. احتجّ ابن الأثير بعدم صحة هذا الحديث في كتاب شرح المسند .

==========

القسم الثالث (٢٠ دقيقة)

املأ الفراغات الآتية بكلمة مناسبة من عندك في ورقة الإجابة

لقد قصد الإسلام أن يكون الإنسان مثلا صالحا محمود الخصال ، شريف الشمائل ، كريم الأخلاق ، إن تكلم صدق ، وإن وعد ـــ٣١ـــ بوعده ، وإن اؤتمن في الأمر ـــ٣٢ـــ الأمانة ولم يخن ، وإن رأى ـــ٣٣ـــ منكرا غيره بيده ، فإن لم ـــ٣٤ـــ فبلسانه ، فإن لم يستطع فبقلبه ، و ـــ٣٥ـــ تكلم خفض صوته ، وإن مشى ـــ٣٦ـــ يكن مختالا فخورا في مشيته ، و ـــ٣٧ـــ رأى كبيرا وقره .

ومن الآداب و ـــ٣٨ـــ السلوك في الإسلام ما يلي :

ـــ٣٩ـــ المسلم أن يحسن الآداب في ـــ٤٠ـــ والمحادثة وأن يتلطف في التخاطب و ـــ٤١ـــ الخشونة في الحديث ، قال تعالى : { وقولوا للناس حسنا } أي كلاما ـــ٤٢ـــ عند المحادثة والمخاطبة فيكون الحديث ـــ٤٣ـــ برفق ليس بالمرتفع ولا بالمنخفض ، و ـــ٤٤ـــ الأمور أوسطها ، ونهى الله عن ـــ٤٥ـــ في الكلام ، قال تعالى : { إن الذين يفترون على الله الكذب لا يفلحون } فالكاذب لا ينجح و ـــ٤٦ـــ يفلح في جميع أموره .

على ـــ٤٧ـــ أن يؤدي التحية الحسنة ويفشي ـــ٤٨ـــ ، قال تعالى : { وإذا حييتم بتحية فحيوا بأحسن منها } . وعلى المسلم أن يوسع لجليسه ـــ٤٩ـــ أقبل عليه ، ويلتزم معه الأدب ـــ٥٠ـــ إذا كان أكبر منه سنا ، و ـــ٥١ـــ إذا كان أبا أو أستاذا ـــ٥٢ـــ .

وليس للقادم أن يقيم أحدا ـــ٥٣ـــ مجلس ليجلس مكانه ، قال صلى الله عليه وسلم (لا يقيم الرجل الرجل من مجلسه ولكن تفسحوا وتوسعوا) .

وعلى المسلم ـــ٥٤ـــ يأكل حتى يجوع ، لأن الأكل ـــ٥٥ـــ الشبع مضرة أكيدة ، والإسلام يراعي ـــ٥٦ـــ الجسد وسلامته . وألا يأكل المسلم ـــ٥٧ـــ يشرب إلا ما أحله الله . قال تعالى : { كلوا من طيبات ما رزقناكم } ـــ٥٨ـــ أراد المسلم أن يأكل فعليه ـــ٥٩ـــ ينظف يديه وفمه ، ثم يسمي بـ ـــ٦٠ـــ الله ، ويبدأ في الأكل بسكينة و ـــ٦١ـــ . وعلى المسلم أن يأكل باليد ـــ٦٢ـــ ومما يليه . وبعد الانتهاء من ـــ٦٣ـــ يحمد المسلم ربه ويشكر على ـــ٦٤ـــ الأكل أو الشرب اقتداء برسول ـــ٦٥ـــ صلى الله عليه وسلم . ثم يغسل يديه وفمه . والإسلام ـــ٦٦ـــ دائما إلى ما فيه الخير ، و ـــ٦٧ـــ للتنظيم ، فقد خصص اليد اليمنى ـــ٦٨ـــ في الأشياء الطيبة الكريمة ، مثل ـــ٦٩ـــ والشرب والمصافحة وحمل المصحف الشريف و ـــ٧٠ـــ العلم ، واليد اليسرى لغير ذلك ، كـ ـــ٧١ـــ وتنظيف الأذن وحمل النعلين . فالشيء ـــ٧٢ـــ يجب أن يستعمل فيه اليد ـــ٧٣ـــ . وكذلك الرِّجْلان فالرجل اليمنى تستعمل لـ ـــ٧٤ـــ في المساجد وعند لبس ـــ ٧٥ـــ ، قال رسول الله صلى الله عليه وسلم :(إذا انتعل أحدكم فليبدأ باليمين وإذا نزع فليبدأ بالشمال).

مع أحر تمنياتي لكم بالتوفيق والنجاح

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : القواعد العربية

الملاحظات :

١. أمامك دفتر للاسئلة وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة .

٣. يحتوي هذا الاختبار على قسمين : لكل قسم تعليمات خاصة للإجابة عن الأسئلة ، اقرأ التعليمات قبل أن تبدأ بالإجابة .

٤. الزمن المخصص للإجابة عن الأسئلة في هذا الاختبار أربعون (٤٠) دقيقة . والزمن المقترح لكل قسم مكتوب في بداية كل قسم .

٥. لكل سؤال درجة واحدة ولا تحاسَب على إجابة خاطئة .

٦. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة الأسئلة .

توقف الآن

لاتفتح ورقة الأسئلة حتى يسمح لك بذلك

القسم الأول (٣٠ دقيقة)

يتكون كل سؤال في هذا القسم من جملة تنقصها كلمة أو عبارة. وبعد كل جملة توجد أربع كلمات أو عبارات . اختر الكلمة أو العبارة التي في رأيك تكون إجابة صحيحة . ثم ضع دائرة حول الحرف الذي يدل على الكلمة أو العبارة التي اخترتها في المكان المخصص في ورقة الإجابة .

١. رأيتُ الطالبات ..... في دراستهن .
ا. يجتهدون
بـ . تجتهدن
جـ . تجتهدين
د. يجتهدن

٢. يسكن في بيت أخي ..... كريم .
ا. رجلان
بـ . رجلا
جـ . رجل
د. رجال

٣. ..... أستاذان بأكاديمية الدراسة الإسلامية وهما عضوان في مجلس الجامعة .
ا. فاطمة وزينب
بـ . إبراهيم وزينب
جـ . خديجة
د. علي ومحمد ويحي

٤. الطلبة والطالبات ..... احتفال العيد الوطني في العاصمة .
ا. يحضرون
بـ . يحضرن
جـ . يحضران
د. تحضران

٥. دخل النصراني في الإسلام فيصبح هو الآن ..... في الدين .
ا. أخينا
بـ . أخانا
جـ . أخونا
د. إخواننا

٦. إن الله يمكّن للمسلمين دينهم ..... ارتضى لهم .
ا. الذين
بـ . التي
جـ . اللذين
د. الذي

٧. تقرأ كلمة "فاطمة" في الجملة : إنّ فاطمة طالبة مجتهدة .بـ :
ا. الكسرة
بـ . الضمة
جـ . الفتحة
د. الفتحتين

٨. ظن سكان القرية أنّ ..... محمد قد غرق في الماء .
ا. أخو
بـ . أخ
جـ . أخي
د. أخا

٩. قلت لنفسي عندما توفي زميلي محمد : يا ليت ....
ا. محمدا موجودا
بـ . محمدا موجود
جـ . محمدين موجودين
د. محمد موجود

١٠ . اسم كان في قوله تعالى : { لقد كان لكم في رسول الله أسوة حسنة ... الآية } هو :
ا. أسوة
بـ . رسول الله
جـ . الله
د. حسنَة

١١. إن الزوج ..... بما يحكم عليه القاضي في مشكلته الزوجية .
ا. راضيا
بـ . راض
جـ . راضيين
د. راضيان

١٢ . أصبح .... ربهم عند الفجر .
ا. المسلمون داعون
بـ . المسلمين داعون
جـ . المسلمون داعين
د. المسلمين داعين

١٣. مثنى كلمة " منتدى " هو :
ا. منتدوان
بـ . منتدان
جـ . منتديان
د. منتدءان

١٤. لم يعلم كثير من الناس أن الإسلام قد ..... الحضارة الراقية .
ا. اجتاز
بـ . جوز
جـ . أجاز
د. تجاوز

١٥. فعل الأمر المفرد المؤنث لفعل " استعاذ " هو :
ا. استعيذي
بـ . استعذي
جـ . تعذي
د. استعوذي

١٦. الأمهات ..... الطعام في المطبخ .
ا. أعدت
بـ . أعد
جـ . أعدوا
د. أعددن

١٧. إن المسلمات .....ليالي رمضان إيمانا واحتسابا .
ا. قامت
بـ . قمن
جـ . قاموا
د. قامتا

١٨. فعل الأمر المفرد المذكر لـ " اتقى " هو :
ا. اتقي
بـ . أق
جـ . وقي
د. اتق

١٩. بدأت عملي ولم ..... منه بعد .
ا. أنتهيا
بـ . أنتهيَ
جـ . أنته
د. أنتهي

٢٠. المسلم والمسلمة لا ..... إلا بالخير دائما .
ا. يدعوا
بـ . يدعوان
جـ . يدعو
د. يدعون

٢١. إننا نحترمك لأنك تأمر المسلمين بالمعروف و ..... عن المنكر .
ا. تنهونه
بـ . تنهانا
جـ . تنهاهم
د. تنهوننا

٢٢. أشفقت على المرأتين ..... فقدتا أبناءهما .
ا. اللتان
بـ . اللواتي
جـ . التي
د. اللتين

٢٣. المبتدأ في الجملة : " في مساجد رجال كثيرون يذكرون الله " هو :
ا. مساجد
بـ . رجال
جـ . الله
د. كثيرون

٢٤. قالت الطالبة : أنا ..... إلى الجامعة غدا .
ا. ذاهب
بـ . ذاهبات
جـ . ذاهبا
د. ذاهبة

٢٥. فرض الله الزكاة على الأغنياء . فأصبح ..... الأموال يدفعون زكاتهم ابتغاء مرضاة الله .
ا. ذوُو
بـ . ذوًا
جـ . ذو
د. ذوِي

٢٦. أيها المؤمنون ..... يدا واحدة في المنشط والمكره .
ا. كونوا
بـ . تكونون
جـ . كوني
د. كن

٢٧. سمعت أن ..... علي قد وصلا من السفر .
ا. أخوان
بـ . أخوين
جـ . أخوي
د. أخوا

٢٨. إن ..... بايعن الرسول وآمنّ وعملن الصالحات سيدخلن الجنة .
ا. اللاتي
بـ . اللتين
جـ . التي
د. اللتان

٢٩. بعدما شرح المعلم ، لا تزال .... في رأيي .
ا. المسالتين غامضتين
بـ . المسالتين غامضتان
جـ . المسالتان غامضتان
د. المسالتان غامضتين

٣٠. مثنى كلمة " عصى " هو :
ا. عصان
بـ . عصوان
جـ . عصين
د. عصيان

٣١. مثنى كلمة " حمراء " هو :
ا. حمراوان
بـ . حمراءان
جـ . حمران
د. حمرايان

٣٢. جذور كلمة " انصرف " هو :
ا. نصر
بـ . نرف
جـ . انصر
د. صرف

٣٣. فعل الأمر المفرد المؤنث لـ " أدارّ " هو :
ا. أري
بـ . دري
جـ . أديري
د. أدري

٣٤. ما إعراب كلمة " أخوك " في الجملة الآتية : كان من بين الفائزين في المسابقة أخوك .

ا. فاعل
بـ . اسم كان
جـ . خبر كان
د. مفعول به

٣٥. إعراب كلمة " راع " في الجملة : " كلكم راع وكلكم مسؤول عن رعيته " هو :
ا. مضاف إليه
بـ . صفة لكلكم
جـ . خبر لكلكم
د. مبتدأ مؤخر

٣٦. يعرف المنافق أن الناس لا ..... ومع ذلك ..... في نفاقه .
ا. يحترمونه ..... يستمرون
بـ. يحترمونه ..... يستمر
جـ. يحترمه ..... يستمرون
د. يحترمه ..... يستمر

٣٧. ..... المسلمون بدين الله الحنيف ولا يتفرقون .
ا. يتمسكون
بـ. تتمسكون
جـ. تتمسك
د. يتمسك

٣٨. استفدت من كتب أستاذي ..... استعرتها منه .
ا. اللتين
بـ. الذي
جـ. التي
د. الذين

٣٩. للقضاة العادلين ..... كريم عند الله .
ا. مقام
بـ. مقامون
جـ. المقام
د. مقامين

٤٠. ما إعراب كلمة " لآيات " في قوله تعالى : { إن في خلق السموات والأرض واختلاف الليل والنهار لآيات لأولي الألباب }
ا. خبر إن
بـ. اسم إن
جـ. الحال
د. التمييز

٤١. عاد إلى البلاد الولدان ..... سافرا إلى مكة .
ا. اللذين
بـ. الذين
جـ. اللذان
د. الذي

٤٢. تُقرأ كلمة " قوية " في الجملة : زينب قصيرة لكنها قوية بـ :
ا. الفتحة الظاهرة
بـ. الضمة الظاهرة
جـ. الفتحة المقدرة
د. الكسرة الظاهرة

٤٣. أرى أن .... من أهم كتب المراجع للفقه الإسلامي .
ا. هذان الكتابان
بـ. هذين الكتابان
جـ. هذين الكتابين
د. هذان الكتابين

٤٤. إن ..... البخاري ومسلم من الكتب المعروفة في علم الحديث .
ا. صحيحان
بـ. صحيحين
جـ. صحيح
د. صحيحي

٤٥. اسم كان في قوله تعالى : { فما كان جواب قومه إلا أن قالوا اقتلوه أو حرقوه ... الآية } هو :
ا. أن قالوا
بـ. قومه
جـ. جواب
د. اقتلوه

٤٦. اسم كان في قوله تعالى : { ما كان على النبي من حرج فيما فرض الله ... الآية } هو
ا. النبي
بـ. حرج
جـ. فرض
د. الله

٤٧. جذور كلمة " اشتاق " هو :
ا. شيق
بـ. شتق
جـ. اشتق
د. اشتا

٤٨. ..... نحن شهر رمضان وستة أيام من شوال .
ا. صام
بـ. صاموا
جـ. صمت
د. صمنا

٤٩. جذور كلمة " اصطبر " هو :
ا. طبر
بـ. صبر
جـ. اصبر
د. صطبر

٥٠. قال الطالب لشيخه : إني ..... قومي ليلا ونهارا .
ا. دعا
بـ. دعا أنا
جـ. دعيت
د. دعوت

القسم الثاني (١٠ دقائق)

اقرأ العبارات الاتية . إذا كانت العبارة في رأيك صحيحة ضع علامة (✓) في ورقة الإجابة وإذا كانت العبارة في رأيك خاطئة ضع علامة (×) في ورقة الإجابة ثم اكتب الإجابة الصحيحة كما في المثالين الآتيين :

المثال الأول :

الجملة الآتية جملة فعلية : ذهب محمد إلى المسجد .
العبارة " الجملة الآتية جملة فعلية " صحيحة ولذلك وضعت علامة (✓) في ورقة الإجابة .

المثال الثاني :

الجملة الآتية جملة فعلية : محمد يراجع درسه .
العبارة " الجملة الآتية جملة فعلية " خاطئة ولذلك وضعت علامة (×) في ورقة الإجابة ثم كتبت العبارة الصحيحة وهي : الجملة الآتية جملة اسمية (الإجابة المختصرة كقولك : جملة اسمية تكون مقبولة كذلك) .

٥١. قال تعالى { ألم نجعل له عينين ولسانا وشفتين } .

إعراب كلمة (عينين) في الآية السابقة مفعول به منصوب بالياء .

٥٢. صيغة الأمر في حالة الإفراد للفعل الرباعي (ساوى) هو (ساوِ) .

٥٣. كلمة (رجل) في الجملة : " في الدار رجل " هي الخبر .

٥٤. تقرأ كلمة (عبرة) في الجملة : كان لنا في قصص الأمم السابقة عبرة عظيمة بالفتحة الظاهرة .

٥٥. إعراب كلمة (هناك) في الجملة : " هناك رجل يبحث عن أخيه " هو الخبر المقدم .

٥٦. تثنية كلمتي (الفتى) و (الذكرى) هي (الفتيان) و (الذكريان) .

٥٧. جذور الكلمة لفعل (اشتق) هو (شتق) .

٥٨. كلمة (وعد) مثال للفعل المعتل وكلمة (شدّ) مثال للفعل المضعف .

٥٩. أحد تصريف فعل (صفى) هو (اصطفى) .

٦٠. جذور الكلمة لفعل (اتخذ) هو (خذ) .

مع أحر تمنياتي لكم بالتوفيق والنجاح

بسم الله الرحمن الرحيم

اختبار تحديد المستوى للغة العربية : المقال

الملاحظات :

١. أمامك دفتر للسؤال وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة .

٣. الزمن المخصص لكتابة المقال في هذا الاختبار ثلاثون دقيقة فقط .

٤. الدرجة الكاملة خمسون درجة .

٥. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة السؤال .

توقف الآن

لاتفتح ورقة السؤال حتى يسمح لك بذلك

اكتب مقالة قصيرة تحت موضوع : " **الالتحاق بالجامعة** " مستعينا بالأفكار الآتية :

- حصولك على نتيجة امتحان الشهادة الثانوية (STPM) أو نتيجة الامتحان الأخير في قسم التأهيلي .
- تقديم الطلب للالتحاق بالجامعة (طلب الاستمارة ، إملاؤها ، إرسالها إلى جهة معينة).
- حصولك على القبول لمواصلة الدراسة في الجامعة (استعدادات للسفر : كتب ، ملابس ، شهادات وغيرها) .
- قدومك إلى الجامعة (أول يوم في الجامعة ، الحياة في الجامعة)

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : الإملاء

ستمع جيدا إلى التوجيهات الآتية

.

هذا اختبار الإملاء . ستستمع إلى قطعة تُقرأ ثلاث مرات . عندما تقرأ القطعة للمرة
ولى استمع فقط ولا تكتب شيئا على ورقة الإجابة . وعندما تقرأ القطعة للمرة الثانية
يك أن تكتبها في ورقة الإجابة . وستقرأ القطعة في هذه المرة بقراءة بطيئة في مقاطع
تسنى لك كتابتها . ثم ستقرأ القطعة للمرة الأخيرة للمراجعة وبإمكانك أن تكتب ما
قده في هذه المرة . وبعد ذلك تخصص دقيقتان للمراجعة الأخيرة .

لملاحظة

عندما تسمع الكلمات مثل فصل و وقف ونقطتان فلا تكتبها وإنما ضع علامتها فقط .

ستعد الآن لنبدأ بالقراءة الأولى . استمع جيدا ولا تكتب شيئا على ورقة الإجابة

قال رسول الله صلى الله عليه وسلم / { من رأى منكم منكرا / فليغيره بيده /
ن لم يستطع فبلسانه / فإن لم يستطع فبقلبه / وذلك أضعف الإيمان }.

نفهم من هذال الحديث / أن النهي عن المنكر / واجبٌ / على كل مسلم ومسلمةٍ /.
هذا النهيُ / يقع في ثلاثِ مراحلَ / : أعلاها / أن يمنع مسلمٌ منكرا بيدهِ / أي بقدرتهِ /
كسر زجاجة الخمرِ / ومنعِ الظَّالمِ / من أن يضرب / أو يؤذيَ المظلومَ / . فإذا لم يستطعِ
مسلمُ / أن يفعل ذلك / لضعفه أو للخطورةِ الَّتي ستقع عليهِ / ، انتقل الأمرُ إلى المرحلة
ثانية / وهي / أن يمنع المنكراتِ بلسانهِ / أي / بوعظه وخطبته وكتاباتهِ / وغيرها منَ
نَّشاطاتِ اللِّسانيةِ / . فإذا لم يقدر كذلكَ / بهذه الطريقةِ / ، انتقل الأمر إلى أدنى
راحلِ / وهي / المنع بالقلبِ / بحيث لا يرضى / عن المنكرِ الَّذي يحدث أمامهُ / .

من هذا الحديث / نستنبط / أنه لا يجوز لمسلمٍ / أن يرى منكرا دون أن يقوم بمنعهِ
. وباهتمام المسلمينَ بهذا الأمر النَّبويّ / فقد ضمنوا لأنفسهم / السَّعادةَ / في الحياة
دنيوية والأخروية / .

القراءة للمرة الثانية للإملاء . استمع جيدا واكتب ما تستمع إليه .

قال رسول الله صلى الله عليه وسلم / (١٠) (نقطتان) من رأى منكم منكرا /(٧)
فليغيره بيده / (٦) فإن لم يستطع فبلسانهِ (١٠) فإن لم يستطع فبقلبهِ / (١٠) وذلك
أضعف الإيمانْ / . (١٠) (وقف)

نفهم من هذا الحديثْ / (٨) أن النهي عن المنكرِ / (٨) واجب على / (٣) كل مسلم
ومسلمةٍ / (٦) . (وقف) وهذا النهيُ / (٣) يقع في ثلاثِ مراحلَ / (١٠) : (نقطتان) أعلاها
/ (٣) أن يمنع مسلم منكرا / (٨) بيده أي بقدرتهِ / (٧) ككسر زجاجة الخمرِ / (٧) ومنع
الظالمِ / (٤) من أن يضرب / (٦) أو يؤذي المظلومَ / (٦) . (وقف) فإذا لم يستطعِ الْمسلمُ /
(١٠) أن يفعل ذلكَ / (٥) لضعفه أو للخطورةِ / (٦) التي ستقع عليهِ / (٦) ، (فصل) انتقلَ
الْأمر إلى الْمرحلة الثانيةِ / (١٣) وهي / (٣) أن يمنع المنكراتِ بلسانهِ / (١٣) أي بوعظه
وخطبتهِ / (٧) وكتاباته وغيرها / (٦) منَ النَّشاطاتِ اللِّسانيةِ / (٧) . (وقف) فإذا لم يقدر
كذلكَ / (٨) بهذه الطَّريقةِ / (٤) انتقل الأمر إلى أدنى المراحلِ / (١٣) وهي / (٣) المنعُ
بالقلبِ / (٧) بحيث لا يرضى عن المنكرِ / (٩) الذي يحدث أمامهُ / (٦) . (وقف) من هذا
الحديثْ / (٦) نستنبط / (٣) أنه لا يجوز لمسلمٍ / (٧) أن يرى منكرا / (٤) دون أن يقوم
بمنعهِ / (٧) . (وقف) وباهتمامِ الْمسلمينَ بهذا الأمرِ النَّبويّ / (١٥) فقد ضمنوا لأنفسهم /
(٦) السعادةَ / (٣) في الحياةِ الدُّنيويةِ والأخرويةْ / (٩) . (وقف)

القراءة للمرة الثالثة : استمع جيدا وراجع ما كتبت من القطعة السابقة
(راجع القطعة المقروءة للمرة الأولى).

(الملاحظة للمشرف : انتظر دقيقتين)

تـــــــ

A.2.3 Placement test 1998/99

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : القراءة والمطالعة

للاحظات :

. أمامك دفتر للاسئلة وورقة منفصلة للإجابة .

. اكتب البيانات المطلوبة في ورقة الإجابة .

. يحتوي هذا الاختبار على ثلاثة أقسام : لكل قسم تعليمات خاصة للإجابة عن الأسئلة ، قرأ التعليمات قبل أن تبدأ بالإجابة .

. الزمن المخصص للإجابة عن الأسئلة في هذا الاختبار خمسون دقيقة . والزمن المحدد كل قسم مكتوب في بداية كل قسم .

. لكل سؤال درجة واحدة ولا تحاسَب على إجابة خاطئة .

. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة الأسئلة .

توقف الآن

لاتفتح ورقة الأسئلة حتى يسمح لك بذلك

القسم الأول (١٠ دقائق)

اقرأ كل فقرة ثم أجب عن الأسئلة التي تليها . ضع دائرة حول الحرف الذي يدل على الإجابة الصحيحة في رأيك في ورقة الإجابة .

سئل أحد الخطباء عن الزمن الذي يحتاج إليه لإعداد خطبة يلقيها لمدة عشر دقائق فأجاب : أسبوعين . فقال السائل : فكيف إذا كانت الخطبة التي ستلقيها تحتاج إلى ساعة من الزمن ، فما الوقت الذي تحتاج إليه لإعدادها؟ فرد الخطيب : أسبوع واحد . فسأل السائل السؤال الثالث : فكيف إذا كانت الخطبة تحتاج إلى ساعتين في إلقائها ، فما الوقت الذي تحتاج إليه لإعدادها؟ فأجاب الخطيب : مثل هذه الخطبة لا تحتاج إلى إعداد ، وأنا على استعداد لإلقائها الآن !!!

١. كم من الوقت يحتاج إليه الخطيب لإعداد خطبة تستغرق ساعتين ؟
   ا. أكثر من أسبوعين .
   ب. أقل من أسبوع .
   جـ. لا يزيد عن ساعتين .
   د. لا يحتاج إلى وقت .

٢. من هذا الحوار نفهم أن الخطيب يحتاج إلى وقت قصير لإعداد
   ا. خطبة قصيرة .
   ب. خطبة طويلة .
   جـ. خطبة قصيرة وطويلة .
   د. خطبة قصيرة وإلقائها .

٣. نفهم من هذا الحوار كذلك أن الخطيب يرتجل ويجيد في
   ا. إعداد خطبة .
   ب. إلقاء خطبة طويلة .
   جـ. إجابة أسئلة .
   د. إلقاء خطبة قصيرة .

اعلم أيها القارئ أن للمدخن " ثلاث فوائد " !! الأولى منها أن شعر المدخن لا يشيب ؛ والثانية الكلب يخاف من المدخن ؛ والثالثة اللص لا يدخل بيت المدخن . وتفسير ذلك أن شعر المدخن لا يشيب لأن المدخن عادة يموت مبكرا بسبب التدخين قبل أن يشيب شعره ! أما الكلب فإنه يخاف من المدخن لأن المدخن يتكئ على العصا عند المشي لمرض أصابه بسبب التدخين فيظن الكلب أن المدخن يريد أن يضربه ! وأما الفائدة الثالثة فإن اللص لا يدخل بيت المدخن لأن المدخن عادة أصابه السعال فلا يستطيع أن ينام فيظن اللص أنه يسهر في الليل !

٤. ماذا يقصد الكاتب من كلمة " فوائد " (الشطر الأول) في القطعة السابقة ؟

ا. منافع .
ب. خسائر .
جـ. مفاسد .
د. مغانم .

٥. ينصحنا الكاتب في هذه القطعة بـ

ا. الإبتعاد عن التدخين لما له من مضار .
ب. الإسراع إلى التدخين لما له من فوائد .
جـ. الهروب من خطورة التشيب والكلب واللص .
د. الابتعاد عن اللص إذا كان مدخنا .

٦. الفكرة الأساسية في هذه القطعة هي أن التدخين

ا. ينفع المدخن بفوائد ثلاثة .
ب. يجعل عمر المدخن طويلا .
جـ. يجعل المدخن لا ينام أبدا في الليل .
د. يضر صحة المدخن .

قال صلى الله عليه وسلم : (ما من مولود إلا يولد على الفطرة فأبواه يهودانه أو ينصرانه أو يمجسانه). فالطفل أمانة عند والديه فإن عوداه الخير نشأ عليه وسعد في الدنيا والآخرة وإن أهملاه وتركاه لقرناء السوء شقي وهلك . فاعتدال الأطفال أو انحرافهم إنما يرجع إلى التنشئة والتربية .

٧. هذه الفقرة تبين لنا أهمية :

ا. تربية الأولاد وتنشئتهم .
ب. قرناء الأولاد واعتقاداتهم .
جـ. سعادة الأولاد في الدنيا والآخرة .
د. اعتدال الأولاد وانحرافهم .

٨. تبين الفقرة أن مستقبل الأطفال يعتمد على :

ا. التربية التي اختارها لهم آباؤهم .
ب. الأعمال التي يقوم بها آباؤهم .
جـ. القرناء الذين يعيشون حولهم .
د. الأمانة التي تقع على آبائهم .

فعلى الآباء أن يستمدوا من الإسلام مناهج التربية الصحيحة وذلك بأن يتخذوا من سيرة رسول الله صلى الله عليه وسلم القدوة الحسنة كما بينه سبحانه وتعالى في قوله { لقد كان لكم في رسول الله أسوة حسنة ...الآية } . وبذلك يكون الآباء أنفسهم قدوة حسنة لأبنائهم في القول والعمل .. وعليهم أن يطعموهم مما رزقهم الله من مال حلال وأن يعلموهم من علوم الدين والدنيا ما يبتغون به طاعة الله ورسوله ويحقق لهم النفع والكسب في حياتهم . إذا اتبع الآباء هذا الأسلوب في تنشئة الأبناء يكونون قد جمعوا بين خيري الدنيا والآخرة وبنوا مجتمعهم على ركائز من الأخلاق والعلم والمال ...

٩. من عناصر التربية التي اقترحها الكاتب في هذه الفقرة هي :

I. اتخاذ سيرة الرسول صلى الله عليه وسلم كقدوة .
II. أن يصبح الآباء نموذجا حسنا لأبنائهم .
III. إطعام الآباء أولادهم طعاما حلالا طيبا .

ا .تعليم الآباء أبناءهم علوم الدين فقط.

ا

. ا ، II

. II ، III ، Iv

. ا ، II ، III

. يرى الكاتب أن الآباء الذين يتخذون الطريقة الصحيحة في تربية الأولاد قد ضمنوا
لادهم :

الغناء والتقدم في العلوم الدنيوية.

. الإبتعاد عن مطالب الدنيا.

.. القدرة على تفريق الدنيا من الآخرة.

. السعادة والنجاح في الدنيا والآخرة.

==========

قسم الثاني (٢٠ دقيقة)

اقرأ كل فقرة ثم ضع علامة (✓) أمام عبارة صحيحة وعلامة (X) أمام عبارة خاطئة
ي ورقة الإجابة.

من المعلوم أن الصلاة إذا أديت كلها في الوقت المخصص لها فهي أداء ، وإن فعلت
رة ثانية في الوقت لخلل غير الفساد فهي إعادة ، وإن فعلت بعد الوقت فهي قضاء ،
القضاء : فعل الواجب بعد وقته أما إن أدرك المصلي جزءا من الصلاة في الوقت فهل
قع أداء؟ للفقهاء رأيان : الأول للحنفية ، والحنابلة على الراجح ، والثاني للمالكية
الشافعية. (من كتاب الفقه الإسلامي وأدلته جـ ١ : ص ٥١٦).

١. إذا صلينا نفس الصلاة مرتين في الوقت المخصص لها والصلاة الأولى لم تكن باطلة
تسمى الصلاة الثانية إعادة.

١٢. نفهم من القطعة السابقة أن الفقهاء اختلفوا في صلاة المصلي التي لم تقع في وقتها
كاملة.

توجه الخليفة هارون الرشيد إلى المدينة المنورة وأراد أن يستمع إلى حديث من العالم
الفقيه مالك بن أنس. فأرسل رسولا إليه يطلب منه الحضور إليه. فقال الإمام مالك
للرسول : قل لأمير المؤمنين ، إن طالب العلم يذهب إلى العلم ، فأما العلم فلا يسعى إلى
أحد. واقتنع الخليفة وزار العالم الفقيه في داره ، لكنه أمر بإخلاء المكان من الناس.
فرفض مالك وأصر أن يبقى الناس وقال : إذا منعنا العلم عن عامة الناس ، فلا خير فيه
للخاصة. ووافق الرشيد مرة أخرى على رغبة مالك ، وسمح للناس بسماع الحديث معه.

١٣. هذه القصة تدلنا على أن العلم يؤتى ولا يأتي إلى طالبه.

١٤. أراد الخليفة ألا يجلس الناس معه في مجلس العلم بسبب ارتفاع مكانته أمام الناس.

١٥. من هذه القصة نفهم كذلك أن الإمام مالك لا يحترم الخليفة عندما يرفض طلبه.

كانت الأم لا تقرأ ولا تكتب. هربت في طفولتها من المدرسة لأنها كرهت مبادئ
القراءة والكتابة. وبقيت معها هذه الكراهية حتى تزوجت. ولم تكن تشعر ، بسبب هذا
الجهل ، بأي نقص في حياتها. فإذا أرادت أن تسمع الأخبار فتحت المذياع ، وإذا أرادت
محاسبة أحد استعانت بزوجها.

ثم أنجبت هذه الزوجة غير المتعلمة طفلة. وكبرت الطفلة ودخلت المدرسة. وفي
أحد الأيام أحضرت الطفلة كراستها إلى أمها وطلبت منها أن تساعدها في تهجية إحدى
الكلمات. وخجلت الأم أن تعترف لابنتها بأنها لا تقرأ ولا تكتب. كانت تحب ابنتها حبا
لم تستطع فيه أن تواجهها بالحقيقة المرة. فذهبت على الفور إلى مدرسة ليلية وبدأت
تتعلم القراءة والكتابة. كانت تسهر الليل لتلحق بدروس طفلتها.

وبدأت الأم تساعد ابنتها في مراجعة دروسها عاما بعد عام. واستمرت تتعلم دون
علم ابنتها اثنتي عشرة سنة كاملة. وعندما جاء موعد امتحانات شهادة الدراسة
الثانوية فوجئت الابنة بأن جارتها في امتحان الشهادة هي أمها ، ودهشت الابنة فقد

كانت تتصور أن أمها حصلت على هذه الشادة منذ اثني عشر عاما .

١٦. كانت الأم تحب الدراسة أيام طفولتها وأصبحت ذكية بعد أن تزوجت .

١٧. التحقت الأم بمدرسة ليلية لأن زوجها يريدها أن تعلّم بنتها في البيت .

١٨. درست الأم في مدرسة ليلية مقررات تختلف عن المقررات التي درستها بنتها في المدرسة .

١٩. تمكنت الأم من مساعدة ابنتها بعدما التحقت بمدرسة ليلية .

٢٠. من العبر في هذه القصة هي أن الرغبة القوية تستطيع أن تدفع الانسان إلى العمل

لست أريد من المحافظة على الوقت أن يملأ الوقت كله بالعمل ، وأن تكون الحياة كلها عملا لا راحة فيها وأن تكون عابسة لا ضحك فيها . فقد كان هذا للأسف هو المثل الأعلى في القرون الوسطى ، وكان خير الناس في ذلك الوقت من جد ولم يلعب ، وعبس ولم يضحك واستحضر الموت في كل لحظة . فلم تدخل السعادة قلبه ، ورآه الناس حزينا دائما كأنه راجع من زيارة الجنازة . وكان من خير ما دعا إليه العلماء في هذا العصر الحديث السرور والضحك واللعب في معقول من الوقت ، فذلك ينفع الناس أكثر من الجد الدائم .

أريد ألا تكون أوقات الفراغ طاغية على أوقات العمل ، وألا تكون أوقات الفراغ هي صميم الحياة ، وأوقات العمل على هامشها ، بل أريد أكثر من ذلك أن تكون أوقات الفراغ خاضعة لحكم العقل كأوقات العمل ، فإن كنا في العمل نعمل للغاية والهدف ، فيجب أن نصرف أوقات الفراغ لغاية كذلك ، إما لفائدة صحية كالألعاب الرياضية ، وإما للذة نفسية كالقراءات العلمية والأدبية .

أما أن تكون الغاية هي قتل الوقت ، فليست غاية مشروعة لأن الوقت هو الحياة فقتل الوقت قتل الحياة . فالذين يصرفون أوقاتهم الطويلة في لعبة شطرنج لا يعملون لغاية يرضاها العقل ، وكذلك الذين يتجولون بين المقاهي والأندية والطرقات لا يطلبون

إلا قتل الوقت فإنهم أعداء للوقت .

٢١. اقترح الكاتب في الفقرة الأولى بأن نركز على العمل فقط في حياتنا.

٢٢. رأى الكاتب في الفقرة الأولى كذلك أن الناس في القرون الوسطى يقسمون أوقاتهم لحياتهم تقسيما خاطئا.

٢٣. اقترح لنا الكاتب أننا لا نحتاج إلى وضع الهدف للنشاطات في أوقات الفراغ وذلك لأننا قد وضعناه في أعمالنا .

٢٤. ما وافق الكاتب الذين يجعلون أوقات فراغهم أهمّ من أوقات أعمالهم .

٢٥. اعتبرت لعبة شطرنج والتجول وغيرهما من الأعمال المشروعة شريطة أن تكون هذه الأعمال غرضها قتل الوقت .

عن أبي هريرة رضي الله عنه قال : (سأل رجل رسول الله صلى الله عليه وسلم فقال : يا رسول الله إنا نركب البحر ونحمل معنا القليل من الماء فإن توضأنا به عطشنا أفنتوضأ بماء البحر فقال رسول الله صلى الله عليه وسلم : هو الطهور ماؤه الحل ميتته) رواه الخمسة .

الحديث أخرجه ابن خزيمة وابن حبان في صحيحيهما وابن الجارود في المنتقى والحاكم في المستدرك والدارقطني والبيهقي في سننهما وابن أبي شيبة . وحكى الترمذي عن البخاري تصحيحه وتعقّبه ابن عبد البر بأنه لو كان صحيحا عنده لأخرجه في صحيحه . ثم حكم ابن عبد البر مع ذلك بصحته لتلقي العلماء له بالقبول فرده من حيث الاسناد وقبله من حيث المعنى . وصححه أيضا ابن المنذر وابن منده والبغوي وقال هذا الحديث صحيح متّفق على صحته . وقال ابن الأثير في شرح المسند هذا حديث صحيح مشهور أخرجه الأئمة في كتبهم واحتجوا به ورجاله ثقات ... (مقتطف من كتاب نيل الأوطار للشوكاني بتصرف ، ص ١٧٠)

٢٦. يدور الحديث السابق حول جواز الوضوء بماء البحر سواء أكان ماء الطهور متوفرا م غير متوفر.

٢٧. الحديث المذكور أعلاه أخرجه الدارقطني والبيهقي في سننهما والترمذي في صحيحه.

٢٨. نفهم من القطعة السابقة أن الحديث المذكور لا نجده في صحيح البخاري.

٢٩. الذي يرفض الحديث السابق من حيث الإسناد ويقبله من حيث المعنى هو ابن عبد البر.

٣٠. احتجّ ابن الأثير بعدم صحة هذا الحديث في كتاب شرح المسند.

==========

القسم الثالث (٢٠ دقيقة)

املأ الفراغات الآتية بكلمة مناسبة من عندك في ورقة الإجابة

لقد قصد الإسلام أن يكون الإنسان مثلا صالحا محمود الخصال، شريف الشمائل، كريم الأخلاق، إن تكلم صدق، وإن وعد ـــ٣١ـــ بوعده، وإن اؤتمن في الأمر ـــ٣٢ـــ الأمانة ولم يخن، وإن رأى ـــ٣٣ـــ منكرا غيره بيده، فإن لم ـــ٣٤ـــ فبلسانه، فإن لم يستطع فبقلبه، و ـــ٣٥ـــ تكلم خفض صوته، وإن مشى ـــ٣٦ـــ يكن مختالا فخورا في مشيته، و ـــ٣٧ـــ رأى كبيرا وقره.

ومن الآداب و ـــ٣٨ـــ السلوك في الإسلام ما يلي:
ـــ٣٩ـــ المسلم أن يحسن الآداب في ـــ٤٠ـــ والمحادثة وأن يتلطف في التخاطب و ـــ٤١ـــ الخشونة في الحديث، قال تعالى: {وقولوا للناس حسنا} أي كلاما ـــ٤٢ـــ عند المحادثة والمخاطبة فيكون الحديث ـــ٤٣ـــ ليس بالمرتفع ولا بالمنخفض، و ـــ٤٤ـــ الأمور أوسطها، ونهى الله عن ـــ٤٥ـــ في الكلام، قال تعالى: {إن الذين يفترون على الله الكذب لا يفلحون} فالكاذب لا ينجح و ـــ٤٦ـــ يفلح في جميع أموره.

على ـــ٤٧ـــ أن يؤدي التحية الحسنة ويفشي ـــ٤٨ـــ، قال تعالى: {وإذا حييتم بتحية فحيوا بأحسن منها}. وعلى المسلم أن يوسع لجليسه ـــ٤٩ـــ أقبل عليه، ويلتزم معه الأدب ـــ٥٠ـــ إذا كان أكبر منه سنا، و ـــ٥١ـــ إذا كان أبا أو أستاذا ـــ٥٢ـــ. وليس للقادم أن يقيم أحدا ـــ٥٣ـــ مجلس ليجلس مكانه، قال صلى الله عليه وسلم:
(لا يقيم الرجل الرجل من مجلسه ولكن تفسحوا وتوسعوا).

وعلى المسلم ـــ٥٤ـــ يأكل حتى يجوع، لأن الأكل ـــ٥٥ـــ الشبع مضرة أكيدة، والإسلام يراعي ـــ٥٦ـــ الجسد وسلامته. وألا يأكل المسلم ـــ٥٧ـــ يشرب إلا ما أحله الله. قال تعالى: {كلوا من طيبات ما رزقناكم} ـــ٥٨ـــ أراد المسلم أن يأكل فعليه ـــ٥٩ـــ ينظف يديه وفمه، ثم يسمي ب ـــ٦٠ـــ الله، ويبدأ في الأكل بسكينة و ـــ٦١ـــ. وعلى المسلم أن يأكل باليد ـــ٦٢ـــ ومما يليه. وبعد الانتهاء من ـــ٦٣ـــ يحمد المسلم ربه ويشكر على ـــ٦٤ـــ الأكل أو الشرب اقتداء برسول ـــ٦٥ـــ صلى الله عليه وسلم. ثم يغسل يديه وفمه. والإسلام ـــ٦٦ـــ دائما إلى ما فيه الخير، و ـــ٦٧ـــ للتنظيم، فقد خصص اليد اليمنى ـــ٦٨ـــ في الأشياء الطيبة الكريمة، مثل ـــ٦٩ـــ والشرب والمصافحة وحمل المصحف الشريف و ـــ٧٠ـــ العلم، واليد اليسرى لغير ذلك، ك ـــ٧١ـــ وتنظيف الأذن وحمل النعلين. فالشيء ـــ٧٢ـــ يجب أن يستعمل فيه اليد ـــ٧٣ـــ. وكذلك الرِّجْلان فالرجل اليمنى تستعمل ل ـــ٧٤ـــ في المساجد وعند لبس ـــ٧٥ـــ، قال رسول الله صلى الله عليه وسلم: (إذا انتعل أحدكم فليبدأ باليمين وإذا نزع فليبدأ بالشمال).

مع أحرّ تمنياتي لكم بالتوفيق والنجاح

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : القواعد العربية

الملاحظات :

١. أمامك دفتر للاسئلة وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة .

٣. يحتوي هذا الاختبار على قسمين : لكل قسم تعليمات خاصة للإجابة عن الأسئلة ، اقرأ التعليمات قبل أن تبدأ بالإجابة .

٤. الزمن المخصص للإجابة عن الأسئلة في هذا الاختبار أربعون (٤٠) دقيقة . والزمن المقترح لكل قسم مكتوب في بداية كل قسم .

٥. لكل سؤال درجة واحدة ولا تحاسَب على إجابة خاطئة .

٦. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة الأسئلة .

توقف الآن

لاتفتح ورقة الأسئلة حتى يسمح لك بذلك

القسم الأول (٣٠ دقيقة)

يتكون كل سؤال في هذا القسم من جملة تنقصها كلمة أو عبارة وبعد كل جملة توجد أربع كلمات أو عبارات . اختر الكلمة أو العبارة التي في رأيك تكون إجابة صحيحة . ثم ضع دائرة حول الحرف الذي يدل على الكلمة أو العبارة التي اخترتها في المكان المخصص في ورقة الإجابة .

١. رأيتُ الطالبات ..... في دراستهن .
ا. تجتهد
ب . تجتهدن
ج . تجتهدين
د. يجتهدن

٢. ..... أستاذان بأكاديمية الدراسة الإسلامية وهما عضوان في مجلس الجامعة .
ا. فاطمة وزينب
ب . إبراهيم وزينب
ج . سلوى ونجوى
د علي ومحمد ويحي

٣. الطلبة والطالبات ..... احتفال العيد الوطني في العاصمة .
ا. يحضرون
ب . يحضرن
ج . تحضرون
د تحضرن

٤. دخل النصراني في الإسلام فيصبح هو الآن ..... في الدين .
ا. أخينا
ب . أخانا
ج . أخونا
د. إخواننا

٥. إن الله يمكّن للمسلمين دينهم ..... ارتضى لهم .
ا. الذين
ب . التي
ج . اللذين
د الذي

٦. تقرأ كلمة "فاطمة" في الجملة : إنّ فاطمة طالبة مجتهدة ..بـ :
ا. الكسرة
بـ . الضمة
جـ . الفتحة
د الفتحتين

٧. ظن سكان القرية أنّ ..... محمد قد غرق في الماء .
ا. أخو
بـ . أخ
جـ . أخي
د أخا

٨. قلت لنفسي عندما توفي زميلي محمد : يا ليت ....
ا. محمدا موجودا
بـ . محمدا موجود
جـ . محمد موجودا
د. محمد موجود

٩ . اسم كان في قوله تعالى : { لقد كان لكم في رسول الله أسوة حسنة ... الآية }
هو :
ا. أسوة
بـ . رسول الله
جـ . اللهِ
د حسنة

١٠. إن الزوج ..... بما يحكم عليه القاضي في مشكلته الزوجية .
ا. راضيا
بـ . راض
جـ . راضيين
د. راضيان

١١ . أصبح .... ربهم عند الفجر .
ا. المسلمون داعون
بـ . المسلمين داعون
جـ . المسلمون داعين
د المسلمين داعين

١٢. مثنى كلمة " منتدى " هو :
ا. منتدوان
بـ . منتدان
جـ . منتديان
د منتدءان

١٣. لم يعلم كثير من الناس أن الإسلام قد ..... الحضارة الراقية .
ا. اجتاز
بـ . جوز
جـ . أجاز
د. اجتوّز

١٤. فعل الأمر المفرد المؤنث لفعل " استعاذ " هو :
ا. استعيذي
بـ . استعذي
جـ . تعذي
د. استعوذي

١٥. الأمهات ..... الطعام في المطبخ .
ا. أعدت
بـ . أعددتنّ
جـ . أعدوا
د أعددن

١٦. إن المسلمات .....ليالي رمضان إيمانا واحتسابا .
ا. قامت
بـ . قمن
جـ . قاموا
د. قمتن

١٧. فعل الأمر المفرد المذكر لـ " اتقى " هو :
ا. اتقي
بـ . أق
جـ . وقي
د اتق

١٨. بدأت عملي ولم ..... منه بعد .
ا. أنتهيا
ب . أنتهيَ
جـ . أنته
د. أنتهي

١٩. المسلم والمسلمة لا ..... إلا بالخير دائما .
ا. يدعوا
ب . يدعوان
جـ . يدعو
د. يدعون

٢٠. إننا نحترمك لأنك تأمر المسلمين بالمعروف و ..... عن المنكر .
ا. تنهونه
ب . تنهانا
جـ . تنهاهم
د تنهوننا

٢١. أشفقت على المرأتين ..... فقدتا أبناءهما .
ا. اللتان
ب . اللواتي
جـ . التي
د اللتين

٢٢. المبتدأ في الجملة : " في مساجد رجال كثيرون يذكرون الله " هو :
ا. مساجد
ب . رجال
جـ . الله
د. كثيرون

٢٣. قالت الطالبة : أنا ..... إلى الجامعة غدا .
ا. ذاهب
ب . ذاهبات
جـ . ذاهبا
د. ذاهبة

٢٤. فرض الله الزكاة على الأغنياء . فأصبح ..... الأموال يدفعون زكاتهم ابتغاء مرضاة الله .
ا. ذوُو
ب . ذا
جـ . ذو
د. ذوي

٢٥. أيها المؤمنون ..... يدا واحدة في المنشط والمكره .
ا. كونوا
ب . تكونون
جـ . كوني
د. كن

٢٦. سمعت أن ..... علي قد وصلا من السفر .
ا. أخوان
ب . أخوين
جـ .أخوي
د أخوا

٢٧. إن ..... بايعن الرسول وآمنّ وعملن الصالحات سيدخلن الجنة .
ا. اللاتي
ب . اللتين
جـ . التي
د. اللتان

٢٨. بعدما شرح المعلم ، لا تزال .... في رأيي .
ا. المسألتين غامضتين
ب . المسألتين غامضتان
جـ . المسألتان غامضتان
د. المسألتان غامضتين

٢٩. مثنى كلمة " عصى " هو :
ا. عصان
ب . عصوان
جـ . عصين
د. عصيان

٣٠. مثنى كلمة " حمراء " هو :
ا. حمراوان
ب. حمراءان
جـ. حمران
د حمرايان

٣١. فعل الأمر المفرد المؤنث لـ " أدارَ " هو :
ا. أدر
ب. دري
جـ. أديري
د أدري

٣٢. ما إعراب كلمة " أخوك " في الجملة الآتية : كان من بين الفائزين في المسابقة أخوك .

ا. فاعل
ب. اسم كان
جـ. خبر كان
د. مفعول به

٣٣. إعراب كلمة " راع " في الجملة : " كلكم راع وكلكم مسؤول عن رعيته " هو :
ا. مضاف إليه
ب. صفة لكلكم
جـ. خبر لكلكم
د مبتدأ مؤخر

٣٤. يعرف المنافق أن الناس لا ..... ومع ذلك ..... في نفاقه .
ا. يحترمونه ..... يستمرون
ب. يحترمونه ..... يستمر
جـ. يحترمه ..... يستمرون
٦د. يحترمه ..... يستمر

٣٥. ..... المسلمون بدين الله الحنيف ولا يتفرقون .
ا. يتمسكون
ب. تتمسكون
جـ. تتمسك
د. يتمسك

٣٦. استفدت من كتب أستاذي ..... استعرتها منه .
ا. اللتين
ب. الذي
جـ. التي
د. الذين

٣٧. للقضاة العادلين ..... كريم عند الله .
ا. مقام
ب. مقامون
جـ. المقام
د مقامين

٣٨. ما إعراب كلمة " لآيات " في قوله تعالى : { إن في خلق السموات والأرض واختلاف الليل والنهار لآيات لأولي الألباب }
ا. خبر إن
ب. اسم إن
جـ. الحال
د. التمييز

٣٩. تُقرأ كلمة " قوية " في الجملة : زينب قصيرة لكنها قوية بـ :
ا. الفتحة الظاهرة
ب. الضمة الظاهرة
جـ. الفتحة المقدرة
د الكسرة الظاهرة

٤٠. أرى أن .... من أهم كتب المراجع للفقه الإسلامي .
ا. هذان الكتابان
ب. هذين الكتابان
جـ. هذين الكتابين
د. هذان الكتابين

٤١. إن ..... البخاري ومسلم من الكتب المعروفة في علم الحديث .
ا. صحيحان
ب. صحيحين
جـ. صحيح
د صحيحي

٤٢. اسم كان في قوله تعالى : { ما كان على النبي من حرج فيما فرض الله ... الآية } هو
ا. النبي
ب. حرج
جـ. فرض
د. الله

٤٣. ..... نحن شهر رمضان وستة أيام من شوال .
ا. صام
ب. صاموا
جـ. صوموا
د. صمنا

٤٤. جذور كلمة " اصطبر " هو :
ا. طبر
ب. صبر
جـ. اصبر
د. صطبر

٤٥. قال الطالب لشيخه : إني ..... قومي ليلا ونهارا .
ا. دعا
ب. دعا أنا
جـ. دعيت
د. دعوت

القسم الثاني (١٠ دقائق)

اقرأ العبارات الاتية . إذا كانت العبارة في رأيك صحيحة ضع علامة (✓) في ورقة الإجابة وإذا كانت العبارة في رأيك خاطئة ضع علامة (×) في ورقة الإجابة ثم اكتب الإجابة الصحيحة كما في المثالين الآتيين :

المثال الأول :

الجملة الآتية جملة فعلية : ذهب محمد إلى المسجد .
العبارة " الجملة الآتية جملة فعلية " صحيحة ولذلك وضعت علامة (✓) في ورقة الإجابة .

المثال الثاني :

الجملة الآتية جملة فعلية : محمد يراجع درسه .
العبارة " الجملة الآتية جملة فعلية " خاطئة ولذلك وضعت علامة (×) في ورقة الإجابة ثم كتبت العبارة الصحيحة وهي : الجملة الآتية جملة اسمية (الإجابة المختصرة كقولك : جملة اسمية تكون مقبولة كذلك) .

٤٦. صيغة الأمر في حالة الإفراد للفعل الرباعي (ساوى) هو (ساو) .

٤٧. كلمة (رجل) في الجملة : " في الدار رجل " هي الخبر .

٤٨. كلمة (وعد) مثال للفعل المعتل وكلمة (شدّ) مثال للفعل المضعف .

٤٩. أحد تصريف فعل (صفى) هو (اصطفى) .

٥٠. جذور الكلمة لفعل (اتخذ) هو (خذّ) .

مع أحر تمنياتي لكم بالتوفيق والنجاح

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : المقال

الملاحظات :

١. أمامك دفتر للسؤال وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة .

٣. الزمن المخصص لكتابة المقال في هذا الاختبار ثلاثون (٣٠) دقيقة فقط .

٤. الدرجة الكاملة مائة (١٠٠) درجة .

٥. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة السؤال .

توقف الآن

لاتفتح ورقة السؤال حتى يسمح لك بذلك

اكتب مقالة قصيرة تحت موضوع : " **الالتحاق بالجامعة** " مستعينا بالأفكار الآتية :

- حصولك على نتيجة امتحان الشهادة التوجيهية (STPM) أو نتيجة الامتحان الأخير في القسم التأهيلي (pra-akademi).
- تقديم الطلب للالتحاق بالجامعة (طلب الاستمارة ، إملاؤها ، إرسالها إلى جهة معينة).
- حصولك على القبول لمواصلة الدراسة في الجامعة (استعدادات للسفر : كتب ، ملابس ، شهادات وغيرها) .
- قدومك إلى الجامعة (أول يوم في الجامعة ، الحياة في الجامعة) .

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : الإملاء

استمع جيدا إلى التوجيهات الآتية :

هذا اختبار الإملاء . ستستمع إلى قطعة تُقرأ ثلاث مرات . عندما تقرأ القطعة للمرة الأولى استمع فقط ولا تكتب شيئا على ورقة الإجابة . وعندما تقرأ القطعة للمرة الثانية عليك أن تكتبها في ورقة الإجابة . وستقرأ القطعة في هذه المرة بقراءة بطيئة في مقاطع ليتسنى لك كتابتها . ثم ستقرأ القطعة للمرة الأخيرة للمراجعة وبإمكانك أن تكتب ما تفقده في هذه المرة . وبعد ذلك تخصص دقيقتان للمراجعة الأخيرة .

الملاحظة :

عندما تسمع الكلمات مثل : فصل و وقف ونقطتان فلا تكتبها وإنما ضع علامتها فقط .

استعد الآن لنبدأ بالقراءة الأولى . استمع جيدا ولا تكتب شيئا على ورقة الإجابة .

قال رسول الله صلى الله عليه وسلم / :{ من رأى منكم منكرا / فليغيره بيده / فإن لم يستطع فبلسانه / فإن لم يستطع فبقلبه / وذلك أضعف الإيمان }.

نفهم من هذال الحديث / أن النهي عن المنكرِ / واجبٌ / على كل مسلم ومسلمةٍ /. وهذا النهيُ / يقع في ثلاثِ مراحلَ / : أعلاها / أن يمنع مسلمٌ منكرا بيدهِ / أي بقدرتهِ / ككسر زجاجة الخمرِ / ومنعِ الظَّالمِ / من أن يضرب / أو يؤذيَ المظلومَ / . فإذا لم يستطعِ الْمسلمُ / أن يفعل ذلك / لضعفه أو للخطورةِ الَّتي ستقع عليهِ / ، انتقل الأمرُ إلى المرحلة الثانيةِ / وهي / أن يمنع المنكراتِ بلسانهِ / أي / بوعظه وخطبته وكتاباتهِ / وغيرِها منَ النَّشاطاتِ اللِّسانيةْ / . فإذا لم يقدر كذلكَ / بهذه الطريقةِ / ، انتقل الأمر إلى أدنى المراحلِ / وهي / اَلمنع بالقلبِ / بحيث لا يرضى / عن المنكرِ الَّذي يحدث أمامهُ / .

من هذا الحديث / نستَنبط / أنه لا يجوز لمسلمٍ / أن يرى منكرا دون أن يقوم بمنعهِ / . وباهتمام الْمسلمينَ بهذا الأمر النّبويّ / فقد ضمنوا لأنفسهم / السَّعادةَ / في الحياة الدنيوية والأخرويةْ / .

القراءة للمرة الثانية للإملاء . استمع جيدا واكتب ما تستمع إليه ،

قال رسول الله صلى الله عليه وسلم / (١٠) :(نقطتان) من رأى منكم منكرا /(٧) فليغيره بيده / (٦) فإن لم يستطع فبلسانهِ (١٠) فإن لم يستطع فبقلبهِ / (١٠) وذلك أضعف الإيمانْ / . (١٠) (وقف)

نفهم من هذا الحديثْ / (٨) أن النهي عن المنكرِ / (٨) واجب على / (٣) كل مسلم ومسلمةٍ / (٦) . (وقف) وهذا النهيُ / (٣) يقع في ثلاثِ مراحلَ / (١٠) : (نقطتان) أعلاها / (٣) أن يمنع مسلم منكرا / (٨) بيده أي بقدرتهِ / (٧) ككسر زجاجة الخمرِ / (٧) ومنع الظالمِ / (٤) من أن يضرب / (٦) أو يؤذي المظلومَ / (٦) . (وقف) فإذا لم يستطعِ الْمسلمُ / (١٠) أن يفعل ذلكَ / (٥) لضعفه أو للخطورةِ / (٦) التي ستقع عليهِ / (٦) ، (فصل) انتقلَ الأمر إلى الْمرحلة الثانيةِ / (١٣) وهي / (٣) أن يمنع المنكراتِ بلسانهِ / (١٣) أي بوعظه وخطبتهِ / (٧) وكتاباته وغيرها / (٦) منَ النَّشاطاتِ اللِّسانيةِ / (٧) . (وقف) فإذا لم يقدر كذلكَ / (٨) بهذه الطَّريقةِ / (٤) انتقل الأمر إلى أدنى المراحلِ / (١٢) وهي / (٣) المنعُ بالقلبِ / (٧) بحيث لا يرضى عن المنكرِ / (٩) الذي يحدث أمامهُ / (٦) . (وقف) من هذا الحديثْ / (٦) نستنبط / (٣) أنه لا يجوز لمسلمٍ / (٧) أن يرى منكرا / (٤) دون أن يقوم بمنعهِ / (٧) . (وقف) وباهتمامِ الْمسلمينَ بهذا الأمرِ النَّبويّ / (١٥) فقد ضمنوا لأنفسهم / (٦) السعادةَ / (٣) في الحياةِ الدُّنيويةِ والأخرويةْ / (٩) . (وقف)

القراءة للمرة الثالثة : استمع جيدا وراجع ما كتبت من القطعة السابقة (راجع القطعة المقروءة للمرة الأولى).

(الملاحظة للمشرف : انتظرْ دقيقتين)

تـــــــ

A.2.4 Placement test 1999/00

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : القراءة والمطالعة

الملاحظات :

١. أمامك دفتر للاسئلة وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة .

٣. يحتوي هذا الاختبار على ثلاثة أقسام : لكل قسم تعليمات خاصة للإجابة عن الأسئلة ، اقرأ التعليمات قبل أن تبدأ بالإجابة .

٤. الزمن المخصص للإجابة عن الأسئلة في هذا الاختبار خمسون دقيقة . والزمن المحدد لكل قسم مكتوب في بداية كل قسم .

٥. لكل سؤال درجة واحدة ولا تحاسَب على إجابة خاطئة .

٦. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة الأسئلة .

توقف الآن

لاتفتح ورقة الأسئلة حتى يسمح لك بذلك

القسم الأول (١٠ دقائق)

اقرأ كل فقرة ثم أجب عن الأسئلة التي تليها . سوّد الحرف الذي يدل على الإجابة الصحيحة في رأيك في ورقة الإجابة .

سئل أحد الخطباء عن الزمن الذي يحتاج إليه لإعداد خطبة يلقيها لمدة عشر دقائق فأجاب : أسبوعين . فقال السائل : فكيف إذا كانت الخطبة التي ستلقيها تحتاج إلى ساعة من الزمن ، فما الوقت الذي تحتاج إليه لإعدادها؟ فرد الخطيب : أسبوع واحد . فسأل السائل السؤال الثالث : فكيف إذا كانت الخطبة تحتاج إلى ساعتين في إلقائها ، فما الوقت الذي تحتاج إليه لإعدادها؟ فأجاب الخطيب : مثل هذه الخطبة لا تحتاج إلى إعداد ، وأنا على استعداد لإلقائها الآن !!!

1. كم من الوقت يحتاج إليه الخطيب لإعداد خطبة تستغرق ساعتين ؟
   A. أكثر من أسبوعين .
   B . أقل من أسبوع .
   C. لا يزيد عن ساعتين .
   D . لا يحتاج إلى وقت .

2. من هذا الحوار نفهم أن الخطيب يحتاج إلى وقت قصير لإعداد
   A. خطبة قصيرة .
   B . خطبة طويلة .
   C. خطبة قصيرة وطويلة .
   D . خطبة قصيرة وإلقائها .

3. نفهم من هذا الحوار كذلك أن الخطيب يرتجل ويجيد في
   A. إعداد خطبة .
   B . إلقاء خطبة طويلة .
   C. إجابة أسئلة .
   D. إلقاء خطبة قصيرة .

اعلم أيها القارئ أن للمدخن " ثلاث فوائد "!! الأولى منها أن شعر المدخن لا يشيب ؛ والثانية الكلب يخاف من المدخن ؛ والثالثة اللص لا يدخل بيت المدخن . وتفسير ذلك أن شعر المدخن لا يشيب لأن المدخن عادة يموت مبكرا بسبب التدخين قبل أن يشيب شعره ! أما الكلب فإنه يخاف من المدخن لأن المدخن يتكئ على العصا عند المشي لمرض أصابه بسبب التدخين فيظن الكلب أن المدخن يريد أن يضربه ! وأما الفائدة الثالثة فإن اللص لا يدخل بيت المدخن لأن المدخن عادة أصابه السعال فلا يستطيع أن ينام فيظن اللص أنه يسهر في الليل !

4. ماذا يقصد الكاتب من كلمة " فوائد " (الشطر الأول) في القطعة السابقة ؟
A. منافع .
B . خسائر .
C. مفاسد .
D . مغانم .

5. ينصحنا الكاتب في هذه القطعة بـ
A. الإبتعاد عن التدخين لما له من مضار .
B. الإسراع إلى التدخين لما له من فوائد .
C. الهروب من خطورة التشيب والكلب واللص .
D. الابتعاد عن اللص إذا كان مدخنا .

6. الفكرة الأساسية في هذه القطعة هي أن التدخين
A. ينفع المدخن بفوائد ثلاثة .
B . يجعل عمر المدخن طويلا .
C . يجعل المدخن لا ينام أبدا في الليل .
D. يضر صحة المدخن .

قال صلى الله عليه وسلم : (**ما من مولود إلا يولد على الفطرة فأبواه يهودانه أو ينصرانه أو يمجسانه**). فالطفل أمانة عند والديه فإن عوداه الخير نشأ عليه وسعد في الدنيا والآخرة وإن أهملاه وتركاه لقرناء السوء شقي وهلك. فاعتدال الأطفال أو انحرافهم إنما يرجع إلى التنشئة والتربية .

7. هذه الفقرة تبين لنا أهمية :
A. تربية الأولاد وتنشئتهم .
B. قرناء الأولاد واعتقاداتهم .
C . سعادة الأولاد في الدنيا والآخرة .
D. اعتدال الأولاد وانحرافهم .

8. تبين الفقرة أن مستقبل الأطفال يعتمد على :
A. التربية التي اختارها لهم آباؤهم .
B. الأعمال التي يقوم بها آباؤهم .
C. القرناء الذين يعيشون حولهم .
D. الأمانة التي تقع على آبائهم .

فعلى الآباء أن يستمدوا من الإسلام مناهج التربية الصحيحة وذلك بأن يتخذوا من سيرة رسول الله صلى الله عليه وسلم القدوة الحسنة كما بينه سبحانه وتعالى في قوله { لقد كان لكم في رسول الله أسوة حسنة ...الآية } . وبذلك يكون الآباء أنفسهم قدوة حسنة لأبنائهم في القول والعمل .. وعليهم أن يطعموهم مما رزقهم الله من مال حلال وأن يعلموهم من علوم الدين والدنيا ما يبتغون به طاعة الله ورسوله ويحقق لهم النفع والكسب في حياتهم . إذا اتبع الآباء هذا الأسلوب في تنشئة الأبناء يكونون قد جمعوا بين خيري الدنيا والآخرة وبنوا مجتمعهم على ركائز من الأخلاق والعلم والمال ...

9. من عناصر التربية التي اقترحها الكاتب في هذه الفقرة هي :
   I. اتخاذ سيرة الرسول صلى الله عليه وسلم كقدوة .
   II. أن يصبح الآباء نموذجا حسنا لأبنائهم .
   III. إطعام الآباء أولادهم طعاما حلالا طيبا .
   Iv .تعليم الآباء أبناءهم علوم الدين فقط .

A. I
B. I ، II
C. II ، III ، Iv
D . I ، II ، III

10. يرى الكاتب أن الآباء الذين يتخذون الطريقة الصحيحة في تربية الأولاد قد ضمنوا لأولادهم :
   A. الغناء والتقدم في العلوم الدنيوية .
   B. الإبتعاد عن مطالب الدنيا .
   C. القدرة على تفريق الدنيا من الآخرة .
   D . السعادة والنجاح في الدنيا والآخرة .

==========

القسم الثاني (٢٠ دقيقة)

اقرأ كل فقرة من الفقرات الآتية ثم أجب عن الاسئلة التي تليها بتسويد حرف (A) أمام عبارة صحيحة وحرف (B) أمام عبارة خاطئة في ورقة الإجابة.

من المعلوم أن الصلاة إذا أديت كلها في الوقت المخصص لها فهي أداء ، وإن فعلت مرة ثانية في الوقت لخلل غير الفساد فهي إعادة ، وإن فعلت بعد الوقت فهي قضاء ، والقضاء : فعل الواجب بعد وقته. أما إن أدرك المصلي جزءا من الصلاة في الوقت فهل تقع أداء؟ للفقهاء رأيان : الأول للحنفية ، والحنابلة على الراجح ، والثاني للمالكية والشافعية . (من كتاب الفقه الإسلامي وأدلته جـ ١ : ص ١٦٥) .

11. إذا صلينا نفس الصلاة مرتين في الوقت المخصص لها والصلاة الأولى لم تكن باطلة فتسمى الصلاة الثانية إعادة .
12. نفهم من القطعة السابقة أن الفقهاء اختلفوا في صلاة المصلي التي لم تقع في وقتها كاملة .

توجه الخليفة هارون الرشيد إلى المدينة المنورة وأراد أن يستمع إلى حديث من العالم الفقيه مالك بن أنس. فأرسل رسولا إليه يطلب منه الحضور إليه . فقال الإمام مالك للرسول : قل لأمير المؤمنين ، إن طالب العلم يذهب إلى العلم ، فأما العلم فلا يسعى إلى أحد . واقتنع الخليفة وزار العالم الفقيه في داره ، لكنه أمر بإخلاء المكان من الناس . فرفض مالك وأصر أن يبقى الناس وقال : إذا منعنا العلم عن عامة الناس ، فلا خير فيه للخاصة . ووافق الرشيد مرة أخرى على رغبة مالك ، وسمح للناس بسماع الحديث معه .

13. هذه القصة تدلنا على أن العلم يؤتى ولا يأتي إلى طالبه .
14. أراد الخليفة ألا يجلس الناس معه في مجلس العلم كي لا يعلموا جهله .
15. من هذه القصة نفهم كذلك أن الإمام مالك لا يحترم الخليفة عندما يرفض طلبه .

كانت الأم لا تقرأ ولا تكتب . هربت في طفولتها من المدرسة لأنها كرهت مبادئ القراءة والكتابة . وبقيت معها هذه الكراهية حتى تزوجت . ولم تكن تشعر ، بسبب هذا الجهل ، بأي نقص في حياتها . فإذا أرادت أن تسمع الأخبار فتحت المذياع ، وإذا أرادت محاسبة أحد استعانت بزوجها .

ثم أنجبت هذه الزوجة غير المتعلمة طفلة . وكبرت الطفلة ودخلت المدرسة . وفي أحد الأيام أحضرت الطفلة كراستها إلى أمها وطلبت منها أن تساعدها في تهجية إحدى الكلمات . وخجلت الأم أن تعترف لابنتها بأنها لا تقرأ ولا تكتب . كانت تحب ابنتها حبا لم تستطع فيه أن تواجهها بالحقيقة المرة . فذهبت على الفور إلى مدرسة ليلية وبدأت تتعلم القراءة والكتابة . كانت تسهر الليل لتلحق بدروس طفلتها .

وبدأت الأم تساعد ابنتها في مراجعة دروسها عاما بعد عام . واستمرت تتعلم دون علم ابنتها اثنتي عشرة سنة كاملة . وعندما جاء موعد امتحانات شهادة الدراسة الثانوية فوجئت الابنة بأن جارتها في امتحان الشهادة هي أمها ، ودهشت الابنة فقد كانت تتصور أن أمها حصلت على هذه الشادة منذ اثني عشر عاما .

16. كانت الأم تحب الدراسة أيام طفولتها وأصبحت ذكية بعد أن تزوجت .
17. التحقت الأم بمدرسة ليلية لأن زوجها يريدها أن تعلّم بنتها في البيت .
18. درست الأم في مدرسة ليلية مقررات تختلف عن المقررات التي درستها بنتها في المدرسة .
19. تمكنت الأم من مساعدة ابنتها بعدما حصلت على الشهادة من مدرسة ليلية .
20. من العبر في هذه القصة هي أن الرغبة القوية تستطيع أن تدفع الانسان إلى العمل

لست أريد من المحافظة على الوقت أن يملأ الوقت كله بالعمل ، وأن تكون الحياة كلها عملا لا راحة فيها وأن تكون عابسة لا ضحك فيها . فقد كان هذا للأسف هو المثل الأعلى في القرون الوسطى ، وكان خير الناس في ذلك الوقت من جد ولم يلعب ، وعبس ولم يضحك واستحضر الموت في كل لحظة . فلم تدخل السعادة قلبه ، ورآه الناس حزينا دائما كأنه راجع من زيارة الجنازة . وكان من خير ما دعا إليه العلماء في هذا العصر الحديث السرور والضحك واللعب في معقول من الوقت ، فذلك ينفع الناس أكثر من الجد الدائم .

أريد ألا تكون أوقات الفراغ طاغية على أوقات العمل ، وألا تكون أوقات الفراغ هي صميم الحياة ، وأوقات العمل على هامشها ، بل أريد أكثر من ذلك أن تكون أوقات الفراغ خاضعة لحكم العقل كأوقات العمل ، فإن كنا في العمل نعمل للغاية والهدف ، فيجب أن نصرف أوقات الفراغ لغاية كذلك ، إما لفائدة صحية كالألعاب الرياضية ، وإما للذة نفسية كالقراءات العلمية والأدبية .

أما أن تكون الغاية هي قتل الوقت ، فليست غاية مشروعة لأن الوقت هو الحياة فقتل الوقت قتل الحياة . فالذين يصرفون أوقاتهم الطويلة في لعبة شطرنج لا يعملون لغاية يرضاها العقل ، وكذلك الذين يتجولون بين المقاهي والأندية والطرقات لا يطلبون إلا قتل الوقت فإنهم أعداء للوقت .

21. اقترح الكاتب في الفقرة الأولى بأن نركز على العمل فقط في حياتنا.
22. رأى الكاتب في الفقرة الأولى كذلك أن الناس في القرون الوسطى يقسمون أوقاتهم لحياتهم تقسيما خاطئا.
23. اقترح لنا الكاتب أننا لا نحتاج إلى وضع الهدف للنشاطات في أوقات الفراغ وذلك لأننا قد وضعناه في أعمالنا .
24. ما وافق الكاتب الذين يجعلون أوقات فراغهم أهمّ من أوقات أعمالهم .
25. اعتبرت لعبة شطرنج والتجول وغيرهما من الأعمال المشروعة شريطة أن تكون هذه الأعمال غرضها قتل الوقت .

عن أبي هريرة رضي الله عنه قال : (سأل رجل رسول الله صلى الله عليه وسلم فقال : يا رسول الله إنا نركب البحر ونحمل معنا القليل من الماء فإن توضأنا به عطشنا أفنتوضأ بماء البحر فقال رسول الله صلى الله عليه وسلم : هو الطهور ماؤه الحل ميتته) رواه الخمسة .

الحديث أخرجه ابن خزيمة وابن حبان في صحيحيهما وابن الجارود في المنتقى والحاكم في المستدرك والدارقطني والبيهقي في سننهما وابن أبي شيبة . وحكى الترمذي عن البخاري تصحيحه وتعقّبه ابن عبد البر بأنه لو كان صحيحا عنده لأخرجه في صحيحه . ثم حكم ابن عبد البر مع ذلك بصحته لتلقي العلماء له بالقبول فرده من حيث الاسناد وقبله من حيث المعنى . وصححه أيضا ابن المنذر وابن منده والبغوي وقال هذا الحديث صحيح متفق على صحته . وقال ابن الأثير في شرح المسند هذا حديث صحيح مشهور أخرجه الأئمة في كتبهم واحتجوا به ورجاله ثقات ... (مقتطف من كتاب نيل الأوطار للشوكاني بتصرف ، ص . ١٧)

26. يدور الحديث السابق حول جواز الوضوء بماء البحر سواء أكان ماء الطهور متوفرا أم غير متوفر.

27. الحديث المذكور أعلاه أخرجه الدارقطني والبيهقي في سننهما والترمذي في صحيحه.

28. نفهم من القطعة السابقة أن الحديث المذكور لا نجده في صحيح البخاري .

29. الذي يرفض الحديث السابق من حيث الإسناد ويقبله من حيث المعنى هو ابن عبد البر .

30. احتجّ ابن الأثير بعدم صحة هذا الحديث في كتاب شرح المسند .

==========

القسم الثالث (٢٠ دقيقة)

اختر الإجابة الصحيحة للفراغات الآتية بكلمة مناسبة من القوسين. ثم سوّد الحرف الذي فيه إجابة صحيحة في ورقة الإجابة :

لقد قصد الإسلام أن يكون الإنسان مثلا صالحا محمود الخصال ، شريف الشمائل ، كريم الأخلاق ، إن تكلم صدق ، وإن وعد ـــ31ـــ (A. وفّر B. أوفى C. خالف D. أخبر) بوعده ، وإن اؤتمن في الأمر ـــ32ـــ (A. كتم B. أدّى C. باع D. صدق) الأمانة ولم يخن ، وإن رأى ـــ33ـــ (A. رجلا B. أمرا C. نقودا D. طريقا) منكرا غيره بيده ، فإن لم ـــ34ـــ (A. يستطيع B. يغير C. يخالف D. يستطع) فبلسانه ، فإن لم يستطع فبقلبه ، و ـــ35ـــ (A. إن B. لم C. لولا D. لا) تكلم خفض صوته ، وإن مشى ـــ36ـــ (A. إن B. لما C. لم D. لا) يكن مختالا فخورا في مشيته ، و ـــ37ـــ (A. لولا B. حتى C.إن D. لا) رأى كبيرا وقره .

ومن الآداب و ـــ38ـــ (A. محاسن B. مكاتب C. مخارج D. موارد) السلوك في الإسلام ما يلي :

ـــ39ـــ (A. ب B. من C. إلى D. على) المسلم أن يحسن الآداب في ـــ40ـــ (A. العمل B. الاستماع C. اللعبة D. المشاركة) والمحادثة وأن يتلطف في التخاطب و ـــ41ـــ (A. يرتكب B. يتجنب C. يمارس D. يعمل) الخشونة في الحديث ، قال تعالى : { وقولوا للناس حسنا } أي كلاما ـــ42ـــ (A. خشنا B. ميعا C. هينا D. حيا) عند المحادثة والمخاطبة فيكون الحديث ـــ43ـــ (A. لينا B. ركيكا C. متذبذبا D. موسيقيا) ليس بالمرتفع ولا بالمنخفض ، و ـــ44ـــ (A. أحسن B. أقبح C. أبسط D. أكبر) الأمور أوسطها ، ونهى الله عن ـــ45ـــ (A. التكبر B.المزاح C. التلطف D. الكذب) في الكلام قال تعالى : { إن الذين يفترون على الله الكذب لا يفلحون } فالكاذب لا

ينجح و

ـــ46 (A. سوف B. لا C. لما D. سـ) يفلح في جميع أموره.

على ـــ47 (A. الرجل B. الابن C. التلميذ D. المسلم) أن يؤدي التحية الحسنة ويفشي ـــ48 (A. الاحترام B. التحية C. الدعاء D. السلام). وعلى المسلم أن يوسع لجليسه ـــ49 (A. إذا B. بعد أن C. قبل أن D. لولا) أقبل عليه، ويلتزم معه الأدب و ـــ50 (A. المشاركة B. الجلوس C. الوقار D. القيام) إذا كان أكبر منه سنا، و ـــ51 (A. عجبا B. ضرورة C. خاصة D. فورا) إذا كان أبا أو أستاذا ـــ52 (A. منه B. قبله C. إليه D. له). وليس للقادم أن يقيم أحدا ـــ53 (A. لـ B. من C. إلى D. على) مجلس ليجلس مكانه.

وعلى المسلم ـــ54 (A. أن B. دائما C. ألا D. قليلا) يأكل حتى يجوع، لأن الأكل ـــ55 (A. حتى B. من C. إلى D. قبل) الشبع مضرة أكيدة، والإسلام يراعي ـــ56 (A.صحة B. مضرة C. قوة D. طاقة) الجسد وسلامته. وألا يأكل المسلم و ـــ57 (A. لما B. لم C. لن D. لا) يشرب إلا ما أحله الله. قال تعالى: { **كلوا من طيبات ما رزقناكم** } . ـــ58 (A. لولا B. إذا C. بل D. لكن) أراد المسلم أن يأكل فعليه ـــ59 (A. ألا B. من C. أن D. على) ينظف يديه وفمه، ثم يسمي بـ ـــ60 (A. صفة B. اسم C. قدرة D. عزة) الله، ويبدأ في الأكل بسكينة و ـــ61 (A. سرعة B. عجلة C. بطء D. هدوء). وعلى المسلم أن يأكل باليد ـــ62 (A. اليسرى B. الكبرى C. الوسطى D. اليمنى) ومما يليه. وبعد الانتهاء من ـــ63 (A. التحية B. الغسل C. الأكل D. البسملة) يحمد المسلم ربه ويشكره على ـــ64 (A. نعمة B. تناول C. انتهاء D. طعام) الأكل أو الشرب اقتداء برسول ـــ65 (A. الكريم B. المختار C. الله D. المصطفى) صلى الله عليه وسلم. ثم يغسل يديه وفمه.

والإسلام ـــ66 (A. يتوجه B. يرشد C. يبسط D. يتقاسم) دائما إلى ما فيه الخير، و ـــ67 (A. يميل B. يدعو C. يتقرب D. يشير) للتنظيم، فقد خصص اليد اليمنى ـــ68 (A. في B. من C. إلى D. على) الأشياء الطيبة الكريمة النبيلة مثل ـــ69 (A. العمل B. الضرب C. الكتابة D. الأكل) والشرب والمصافحة وحمل المصحف الشريف و ـــ70 (A. تناول B. أدوات C. مجالات D. كتب) العلم، واليد اليسرى لغير ذلك، كـ ـــ71 (A. السرقة B. الاستنجاء C. اللعب D. ضرب العدو) وتنظيف الأذن وحمل النعلين. فالشيء ـــ72 (A. الطويل B. الكبير C. الحسن D. المتين) يجب أن يستعمل فيه اليد ـــ73 (A. الطويلة B. اليمنى C. الكبرى D. اليسرى). وكذلك الرِّجْلان فالرجل اليمنى تستعمل لـ ـــ74 (A. الدخول B. السجود C. الصلاة D. الخطبة) في المساجد وعند لبس ـــ 75 (A. الثوب B. الجورب C. النعال D. السروال) امتثالا لقول رسول الله صلى الله عليه وسلم: (**إذا انتعل أحدكم فليبدأ باليمين وإذا نزع فليبدأ بالشمال**).

مع أحر تمنياتي لكم بالتوفيق والنجاح

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : القواعد العربية

الملاحظات :

١. أمامك دفتر للاسئلة وورقة منفصلة للإجابة .

٢. اكتب البيانات المطلوبة في ورقة الإجابة واكتب اسمك بالحروف اللاتينية .

٣. يحتوي هذا الاختبار على قسمين : لكل قسم تعليمات خاصة للإجابة عن الأسئلة ، اقرأ التعليمات قبل أن تبدأ بالإجابة .

٤. الزمن المخصص للإجابة عن الأسئلة في هذا الاختبار خمس وثلاثون (٣٥) دقيقة . والزمن المقترح لكل قسم مكتوب في بداية كل قسم .

٥. لكل سؤال درجة واحدة ولا تحاسَب على إجابة خاطئة .

٦. اكتب إجابتك على ورقة الإجابة ولا تكتب شيئا على ورقة الأسئلة .

توقف الآن

لاتفتح ورقة الأسئلة حتى يسمح لك بذلك

القسم الأول (٣٠ دقيقة)

يتكون كل سؤال في هذا القسم من جملة تنقصها كلمة أو عبارة. وبعد كل جملة توجد أربع كلمات أو عبارات . اختر الكلمة أو العبارة التي في رأيك تكون إجابة صحيحة . ثم سوّد الحرف الذي يدل على الكلمة أو العبارة التي اخترتها في المكان المخصص في ورقة الإجابة .

1. أعجبتني الطالبات اللواتي ..... في دراستهن .
A. تجتهد
B . تجتهدن
C . تجتهدين
D. يجتهدن

2. ..... أستاذان بأكاديمية الدراسة الإسلامية وهما عضوان في مجلس الجامعة .
A. فاطمة وزينب
B . إبراهيم وزينب
C . سلوى ونجوى
D. علي ومحمد ويحي

3. الطلبة والطالبات ..... احتفال العيد الوطني في العاصمة .
A. يحضرون
B . يحضرن
C . تحضرون
D. تحضرن

4. دخل النصراني في الإسلام فيصبح هو الآن ..... في الدين .
A. أخينا
B. أخانا
C . أخونا
D. إخواننا

5. إن الله يمكّن للمسلمين دينهم ..... ارتضى لهم .
A. الذين
B . التي
C . اللذين
D. الذي

6. تقرأ كلمة "فاطمة " في الجملة : إنّ فاطمة طالبة مجتهدة ..بـ :
A. الكسرة
B . الضمة
C . الفتحة
D. الفتحتين

7. ظن سكان القرية أنّ ..... محمد قد غرق في الماء .
A. أخو
B . أخ
C . أخي
D. أخا

8. قلت لنفسي عندما توفي زميلي محمد : يا ليت ....
A. محمدا موجودا
B . محمدا موجود
C . محمد موجودا
D. محمد موجود

9 . اسم كان في قوله تعالى : { لقد كان لكم في رسول الله أسوة حسنة ... الآية } هو :
A. أسوة
B . رسول الله
C . اللهِ
D. حسنة

10. إن الزوج ..... بما يحكم عليه القاضي في مشكلته الزوجية .
A. راضيا
B. راض
C. راضيين
D. راضيان

11 . أصبح .... ربهم عند الفجر .
A. المسلمون داعون
B . المسلمين داعون
C . المسلمون داعين
D. المسلمين داعين

12. مثنى كلمة " منتدى " هو :
A. منتدوان
B . منتدان
C. منتديان
D. منتدءان

13. لم يعلم كثير من الناس أن الإسلام قد ..... الحضارة الراقية .
A. اجتاز
B . جوز
C . أجاز
D. اجتوّز

14. فعل الأمر المفرد المؤنث لفعل " استعاذ " هو :
A. استعيذي
B. استعذي
C. تعذي
D. استعوذي

15. الأمهات ..... الطعام في المطبخ .
A. أعدت
B. أعددتنّ
C . أعدوا
D. أعددن

16. إن المسلمات .....ليالي رمضان إيمانا واحتسابا .
A. قامت
B. قمن
C . قاموا

D. قمتن

17. فعل الأمر المفرد المذكر لـ " اتقى " هو :
A. اتقي
B . أق
C . وقي
D. اتق

18. بدأت عملي ولم ..... منه بعد .
A. أنتهيا
B . أنتهيَ
C . أنته
D. أنتهي

19. المسلم والمسلمة لا ..... إلا بالخير دائما .
A. يدعوا
B. يدعوان
C . يدعو
D. يدعون

20. إننا نحترمك لأنك تأمر المسلمين بالمعروف و ..... عن المنكر .
A. تنهونه
B. تنهانا
C . تنهاهم
D. تنهوننا

21. أشفقت على المرأتين ..... فقدتا أبناءهما .
A. اللتان
B. اللواتي
C . التي
D. اللتين

22. المبتدأ في الجملة : " في مساجد رجال كثيرون يذكرون الله " هو :
A. مساجد
B . رجال
C . الله
D. كثيرون

23. قالت الطالبة : أنا ..... إلى الجامعة غدا .
A. داهب
B. داهبات
C . ذاهبا
D. ذاهبة

24. فرض الله الزكاة على الأغنياء . فأصبح ..... الأموال يدفعون زكاتهم ابتغاء مرضاة الله .
A. ذوُو
B. ذا
C . ذو
D. ذوي

25. أيها المؤمنون ..... يدا واحدة في المنشط والمكره .
A. كونوا
B. تكونون
C . كوني
D. كن

26. سمعت أن ..... علي قد وصلا من السفر .
A. أخوان
B . أخوين
C. أخوي
D. أخوا

27. إن ..... بايعن الرسول وآمنّ وعملن الصالحات سيدخلن الجنة .
A. اللاتي
B. اللتين
C. التي
D. اللتان

28. بعدما شرح المعلم، لا تزال .... في رأيي .
A. المسألتين غامضتين
B. المسألتين غامضتان
C. المسألتان غامضتان
D. المسألتان غامضتين

29. مثنى كلمة " عصى " هو :
A. عصان
B. عصوان
C . عصين
D. عصيان

30. مثنى كلمة " حمراء " هو :
A. حمراوان
B. حمراءان
C . حمران
D. حمرايان

31. فعل الأمر المفرد المؤنث لـ " أدارَ " هو :
A. أدر
B . دري
C . أديري
D. أدري

32. ما إعراب كلمة " أخوك " في الجملة الآتية : كان من بين الفائزين في المسابقة أخوك .
A. فاعل
B. اسم كان
C . خبر كان
D. مفعول به

33. إعراب كلمة " راع " في الجملة : " كلكم راع وكلكم مسؤول عن رعيته " هو :
A. مضاف إليه
B. صفة لكلكم
C . خبر لكلكم
D. مبتدأ مؤخر

34. يعرف المنافق أن الناس لا ..... ومع ذلك..... في نفاقه .
A. يحترمونه ..... يستمرون
B. يحترمونه ..... يستمر
C . يحترمه ..... يستمرون
D ٧ يحترمه ..... يستمر

35. ..... المسلمون بدين الله الحنيف ولا يتفرقون .
A. يتمسكون
B . تتمسكون
C . تتمسك
D. يتمسك

36. استفدت من كتب أستاذي ..... استعرتها منه .
A. اللتين
B . الذي
C . التي
D. الذين

37. للقضاة العادلين ..... كريم عند الله .
A. مقام
B . مقامون
C . المقام
D. مقامين

38. ما إعراب كلمة " لآيات " في قوله تعالى : { إن في خلق السموات والأرض واختلاف الليل والنهار لآيات لأولي الألباب }
A. خبر إن
B . اسم إن
C . الحال
D. التمييز

39. تُقرأ كلمة " قوية " في الجملة : زينب قصيرة لكنها قوية بـ :
A. الفتحة الظاهرة
B . الضمة الظاهرة
C . الفتحة المقدرة
D. الكسرة الظاهرة

40. أرى أن .... من أهم كتب المراجع للفقه الإسلامي .
A. هذان الكتابان
B. هذين الكتابان
C. هذين الكتابين
D. هذان الكتابين

41. إن ..... البخاري ومسلم من الكتب المعروفة في علم الحديث .
A. صحيحان
B. صحيحين
C . صحيح
D. صحيحي

42. اسم كان في قوله تعالى : { ما كان على النبي من حرج فيما فرض الله ... الآية } هو
A. النبي
B . حرج
C . فرض
D. الله

43. ..... نحن شهر رمضان وستة أيام من شوال .
A. صام
B . صاموا
C . صوموا
D. صمنا

44. جذور كلمة " اصطبر " هو :
A. طبر
B. صبر
C . اصبر
D. صطبر

45. قال الطالب لشيخه : إني ..... قومي ليلا ونهارا .
A. دعا
B. دعا أنا
C . دعيت
D. دعوت

==========

اقلب الصفحة

القسم الثاني (٥ دقائق)

اقرأ العبارات الاتية . إذا كانت العبارة في رأيك صحيحة سوّد حرف (A) في ورقة الإجابة وإذا كانت العبارة في رأيك خاطئة سوّد حرف (B) في ورقة الإجابة .

46. صيغة الأمر في حالة الإفراد للفعل الرباعي (ساوى) هو (ساو) .

47. كلمة (رجل) في الجملة : " في الدار رجل " هي الخبر .

48. كلمة (وعد) مثال للفعل المعتل وكلمة (شدّ) مثال للفعل المضعف .

49. أحد تصريف فعل (صفى) هو (اصطفى) .

50. جذور الكلمة لفعل (اتخذ) هو (خذّ) .

51. تقرأ كلمة (عبرة) في الجملة : " كان لنا في قصص الأمم السابقة عبرة عظيمة " بالفتحة الظاهرة.

52. الجملة الآتية جملة فعلية : " أن يراجع محمد درسه خير له من أن يشاهد التلفزيون "

53. صيغة الأمر لفعل (اتقى) في حالة الجمع هي (اتقوْا).

54. الجملة الآتية صحيحة : " قرأتُ كتابا الذي اشتريتُه في الأسبوع الماضي ".

55. خبر إنّ في قولنا : " إن المسلمين والمسلمات إخوة يتحابون ويتعاونون فيما بينهم" هو (يتحابون).

مع أحر تمنياتي لكم بالتوفيق والنجاح

A.2.5 Answer sheets

ورقة الإجابة

القراءة والمطالعة

الإسم : _____

_____

الكلية : _____

القسم الأول :

١. ا ب جـ د

٢. ا ب جـ د

٣. ا ب جـ د

٤. ا ب جـ د

٥. ا ب جـ د

٦. ا ب جـ د

٧. ا ب جـ د

٨. ا ب جـ د

٩. ا ب جـ د

١٠. ا ب جـ د

القسم الثاني :

١١. ( )

١٢. ( )

١٣. ( )

١٤. ( )

١٥. ( )

١٦. ( )

١٧. ( )

١٨. ( )

١٩. ( )

٢٠. ( )

٢١. ( )

٢٢. ( )

٢٣. ( )

٢٤. ( )

٢٥. ( )

٢٦. ( )

٢٧. ( )

٢٨. ( )

٢٩. ( )

٣٠. ( )

القسم الثالث :

٣١. _____

٣٢. _____

٣٣. _____

٣٤. _____

٣٥. _____

٣٦. _____

٣٧. _____

٣٨. _____

٣٩. _____

٤٠. _____

٤١. _____

٤٢. _____

٤٣. _____

٤٤. _____

٤٥. _____

٤٦. _____

٤٧. _____

٤٨. _____

٤٩. _____

٥٠. _____

٥١. _____

٥٢. _____

٥٣. _____

٥٤. _____

٥٥. _____

٥٦. _____

٥٧. _____

٥٨. _____

٥٩. _____

٦٠. _____

٦١. _____

٦٢. _____

٦٣. _____

٦٤. _____

٦٥. _____

٦٦. _____

٦٧. _____

٦٨. _____

٦٩. _____

٧٠. _____

٧١. _____

٧٢. _____

٧٣. _____

٧٤. _____

٧٥. _____

٧٦. _____

٧٧. _____

٧٨. _____

٧٩. _____

٨٠. _____

٨١. _____

٨٢. _____

٨٣. _____

٨٤. _____

٨٥. _____

٨٦. _____

٨٧. _____

٨٨. _____

٨٩. _____

٩٠. _____

٩١. _____

٩٢. _____

٩٣. _____

٩٤. _____

٩٥. _____

٩٦. _____

٩٧. _____

٩٨. _____

٩٩. _____

رقة الإجابة

لقراءة والمطالعة

سم : _____

_____

كلية : _____

درسة الثانوية الأخيرة :

_____

_____

سم الأول :

ا ب جـ د

ا ب جـ د

ا ب جـ د

ا ب جـ د

ا ب جـ د

ا ب جـ د

ا ب جـ د

ا ب جـ د

ا ب جـ د

ا ب جـ د

القسم الثاني :

١١. ( )

١٢. ( )

١٣. ( )

١٤. ( )

١٥. ( )

١٦. ( )

١٧. ( )

١٨. ( )

١٩. ( )

٢٠. ( )

٢١. ( )

٢٢. ( )

٢٣. ( )

٢٤. ( )

٢٥. ( )

٢٦. ( )

٢٧. ( )

٢٨. ( )

٢٩. ( )

٣٠. ( )

القسم الثالث :

٣١. _____

٣٢. _____

٣٣. _____

٣٤. _____

٣٥. _____

٣٦. _____

٣٧. _____

٣٨. _____

٣٩. _____

٤٠. _____

٤١. _____

٤٢. _____

٤٣. _____

٤٤. _____

٤٥. _____

٤٦. _____

٤٧. _____

٤٨. _____

٤٩. _____

٥٠. _____

٥١. _____

٥٢. _____

٥٣. _____

٥٤. _____

٥٥. _____

٥٦. _____

٥٧. _____

٥٨. _____

٥٩. _____

٦٠. _____

٦١. _____

٦٢. _____

٦٣. _____

٦٤. _____

٦٥. _____

٦٦. _____

٦٧. _____

٦٨. _____

٦٩. _____

٧٠. _____

٧١. _____

٧٢. _____

٧٣. _____

٧٤. _____

٧٥. _____

ورقة الإجابة

القواعد العربية

الإسم : _____

_____

الكلية : _____

القسم الأول :

١. ا بـ جـ د
٢. ا بـ جـ د
٣. ا بـ جـ د
٤. ا بـ جـ د
٥. ا بـ جـ د
٦. ا بـ جـ د
٧. ا بـ جـ د
٨. ا بـ جـ د
٩. ا بـ جـ د
١٠. ا بـ جـ د
١١. ا بـ جـ د
١٢. ا بـ جـ د
١٣. ا بـ جـ د
١٤. ا بـ جـ د
١٥. ا بـ جـ د
١٦. ا بـ جـ د
١٧. ا بـ جـ د
١٨. ا بـ جـ د
١٩. ا بـ جـ د
٢٠. ا بـ جـ د
٢١. ا بـ جـ د
٢٢. ا بـ جـ د
٢٣. ا بـ جـ د
٢٤. ا بـ جـ د
٢٥. ا بـ جـ د
٢٦. ا بـ جـ د
٢٧. ا بـ جـ د
٢٨. ا بـ جـ د
٢٩. ا بـ جـ د
٣٠. ا بـ جـ د
٣١. ا بـ جـ د
٣٢. ا بـ جـ د
٣٣. ا بـ جـ د
٣٤. ا بـ جـ د
٣٥. ا بـ جـ د
٣٦. ا بـ جـ د
٣٧. ا بـ جـ د
٣٨. ا بـ جـ د
٣٩. ا بـ جـ د
٤٠. ا بـ جـ د
٤١. ا بـ جـ د
٤٢. ا بـ جـ د
٤٣. ا بـ جـ د
٤٤. ا بـ جـ د
٤٥. ا بـ جـ د
٤٦. ا بـ جـ د
٤٧. ا بـ جـ د
٤٨. ا بـ جـ د
٤٩. ا بـ جـ د
٥٠. ا بـ جـ د

القسم الثاني :

المثال الأول :

(✓)

_____

المثال الثاني:

(X)

<u>الجملة الآتية جملة فعلية</u>

٥١. ( )

_____

٥٢. ( )

_____

٥٣. ( )

_____

٥٤. ( )

_____

٥٥. ( )

_____

٥٦. ( )

_____

٥٧. ( )

_____

٥٨. ( )

_____

٥٩. ( )

_____

٦٠. ( )

_____

٦١. ( )

_____

٦٢. ( )

_____

٦٣. ( )

_____

٦٤. ( )

_____

٦٥. ( )

_____

ورقة الإجابة

القواعد العربية

الاسم : _____

_____

الكلية : _____

القسم الأول :

1 ا بـ جـ د
2 ا بـ جـ د
3 ا بـ جـ د
4 ا بـ جـ د
5 ا بـ جـ د
6 ا بـ جـ د
7 ا بـ جـ د
8 ا بـ جـ د
9 ا بـ جـ د
10 ا بـ جـ د
11 ا بـ جـ د
12 ا بـ جـ د
13 ا بـ جـ د
14 ا بـ جـ د
15 ا بـ جـ د
16 ا بـ جـ د
17 ا بـ جـ د
18 ا بـ جـ د
19 ا بـ جـ د
20 ا بـ جـ د
21 ا بـ جـ د
22 ا بـ جـ د
23 ا بـ جـ د
24 ا بـ جـ د
25 ا بـ جـ د
26 ا بـ جـ د
27 ا بـ جـ د
28 ا بـ جـ د
29 ا بـ جـ د
30 ا بـ جـ د
21 ا بـ جـ د
22 ا بـ جـ د
23 ا بـ جـ د
24 ا بـ جـ د
25 ا بـ جـ د
26 ا بـ جـ د
27 ا بـ جـ د
28 ا بـ جـ د
29 ا بـ جـ د
30 ا بـ جـ د
31 ا بـ جـ د
32 ا بـ جـ د
33 ا بـ جـ د
34 ا بـ جـ د
35 ا بـ جـ د
36 ا بـ جـ د
37 ا بـ جـ د
38 ا بـ جـ د
39 ا بـ جـ د
40 ا بـ جـ د
41 ا بـ جـ د
42 ا بـ جـ د
43 ا بـ جـ د
44 ا بـ جـ د
45 ا بـ جـ د

القسم الثاني :

المثال الأول :

{ }

المثال الثاني :

{ }

الجملة الآتية جملة اسمية

---

46 { }

---

47 { }

---

48 { }

---

49 { }

---

50 { }

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : الإملاء

ورقة الإجابة

الاسم : _____

الكلية : _____

بسم الله الرحمن الرحيم

اختبار تحديد المستوى في اللغة العربية : المقال

ورقة الإجابة

الاسم : _____

الكلية : _____ الزمن : ٣٠ دقيقة

قال رسول الله صلى الله عليه وسلم :{ من رأى منكم منكرا فليغيره بيده
(1)

فإن لم يستطع فبلسانه فإن لم يستطع فبقلبه وذلك أضعف الإيمان }.
(2) (3)

نفهم من هذا الحديث أن النهي عن المنكرِ واجبٌ على كل مسلم ومسلمةٍ .
(4) (5)

وهذا النهيُ يقع في ثلاثِ مراحلَ : أعلاها أن يمنع مسلمٌ منكرا بيدهِ أي بقدرتهِ
(6) (7)

ككسر زجاجة الخمرِ ومنعِ الظَّالمِ من أن يضرب أو يؤذيَ المظلومَ . فإذا لم يستطعِ
(8) (9) (10) (11) (12) (13)

الْمسلمُ أن يفعل ذلك لضعفه أو للخطورة الَّتي ستقع عليه ، انتقل الأمرُ إلى
(14) (15) (16)

المرحلة الثانيةِ وهي أن يمنع المنكَّراتِ بلسانهِ أي بوعظه وخطبته وكتاباته
(17) (18)

وغيرِها منَ النَّشاطاتِ اللِّسانيةْ . فإذا لم يقدر كذلكَ ببذه الطريقةِ ، انتقل
(19)

الأمر إلى أدنى المراحلِ وهي المنع بالقلبِ بحيث لا يرضى عن المنكرِ الَّذي يحدث

أمامهُ .
(20) (21)

من هذا الحديث نستنبط أنه لا يجوز لمسلمٍ أن يرى منكرا دون أن يقوم
(22)

بمنعهِ . وباهتمام الْمسلمينَ بهذا الأمر النَّبويَ فقد ضمنوا لأنفسهم السَّعادةَ في
(23) (24) (25) (26)

الحياة الدنيويةِ والأخرويةْ .
(27)

قال رسول الله صلى الله عليه وسلم :{ من رأى منكم منكرا فليغيره بيده
(1)

فإن لم يستطع فبلسانه فإن لم يستطع فبقلبه وذلك أضعف الإيمان }.
(2) (3)

نفهم من هذا الحديث أن النهي عن المنكرِ واجبٌ على كل مسلم ومسلمةٍ .
(4) (5)

وهذا النهيُ يقع في ثلاثِ مراحلَ : أعلاها أن يمنع مسلمٌ منكرا بيدهِ أي بقدرتهِ
(6) (7)

ككسر زجاجة الخمرِ ومنعِ الظَّالمِ من أن يضرب أو يؤذيَ المظلومَ . فإذا لم يستطعِ
(8) (9) (10) (11) (12)

الْمسلمُ أن يفعل ذلك لضعفه أو للخطورة الَّتي ستقع عليهِ ، انتقل الأمرُ إلى
(13) (14) (15)

المرحلة الثانيةِ وهي أن يمنع المنكراتِ بلسانهِ أي بوعظه وخطبته وكتاباته
(16)

وغيرِها منَ النَّشاطاتِ اللِّسانيةْ . فإذا لم يقدر كذلكَ بهذه الطريقةِ ، انتقل
(17)

الأمر إلى أدنى المراحلِ وهي اَلمنع بالقلبِ بحيث لا يرضى عن المنكرِ الَّذي يحدث

أمامهُ .
(18) (19)

من هذا الحديث نستنبط أنه لا يجوز لمسلمٍ أن يرى منكرا دون أن يقوم
(20)

بمنعهِ . وباهتمام الْمسلمينَ بهذا الأمر النّبويَ فقد ضمنوا لأنفسهم السَّعادةَ في
(21) (22) (23) (24)

الحياة الدنيويةِ والأخرويةْ .
(25)

A.2.5 Final Examination paper for Arabic at the AIS

UNIVERSITI MALAYA

PEPERIKSAAN SEMESTER I

IJAZAH SARJANA MUDA PENDIDIKAN ISLAM

IJAZAH SARJANA MUDA SYARIAH

IJAZAH SARJANA MUDA USULUDDIN

SESI 1998/99

IXEX 1201: BAHASA ARAB I

SEPTEMBER/OKTOBER 1998 MASA : 2 ½ JAM

Perhatian:

1. Jawab semua soalan.
2. Jawapan bagi soalan objektif perlu dijawab di atas kertas jawapan OMR.
3. Calon-calon tidak dibenarkan membawa keluar kertas soalan keluar dari dewan peperiksaan.

ملحوظة:

١ . أجب عن الأسئلة كلها.

٢ . الإجابة عن الأسئلة الموضوعية لا بد أن تكون على ورقة الخاصة للحاسب الآلي.

٣ . لا يسمح لأي طالب بإخراج الأسئلة كلها أو بعضها خارج قاعة الامتحان.

. . . 2/-

(Kertas soalan ini mengandungi 3 bahagian dalam 17 halaman yang dicetak)

القسم الأول : العلوم العربية (النحو والصرف والبلاغة)

١ . ما نوع التنوين في قولنا : {سلمت علىعمرويه وعمرويهٍ آخر}

A . تنوين المقابلة

B . تنوين التمكين

C . تنوين العوض عن كلمة

D . تنوين التنكير

E . تنوين العوض

٢. سبب امتناع تنوين الكلمة التي تحتها خط في العبارة الآتية:

{جاء أحمدُ إلى المحاضرة مبكرا}

A . فاعل مرفوع

B . مضاف

C . مفرد مذكر

D . ممنوع من الصرف

E . علم موصوف

٣. عين عنصري الجملة الآتية : {المتوازيان لا يلتقيان}

A . فاعل – فعل B . اسم – فعل C . مبتدأ – خبر

D . خبر مقدم – مبتدأ E . خبر – مبتدأ

٤. متى يدل الفعل الماضي على الاستقبال ؟

A . إذا وقع بعد "إذا" أو "إن"

B . إذا أريد به الإنشاء

C . إذا دل على الماضي

D . إذا وقع بعد "ما"

E . إذا أريد به الدعاء

٥. التاء في {كتبْتُ الرسالة}

A . ساكنة

B . متحركة

C . ضمير مستتر

D . فاعل مرفوع

E . ضمير منفصل

٦. علامة الاسم لكلمة "الرحيم" في قوله تعالى {بسم الله الرحمن الرحيم}

A . الإسناد B . الجر بالإضافة C . الجر بالحرف

D . الوصف E . الجر بالتبعية

٧. كل ما يأتي علامات الفعل إلا:

A . تاء الفاعل

B . تاء التأنيث الساكنة

C . ياء المخاطبة

D . نون التوكيد

E . تاء التأنيث المتحركة

٨. علامة نصب الأسماء الخمسة:

A . حذف حرف العلة

B . الياء

C . حذف النون

D . الألف

E . الفتحة

٩. إعراب كلمة "الفتى" في العبارة التالية {رأيت الفتى يؤدي واجبه}

A . فاعل مرفوع بالضمة المقدرة

B . فاعل مرفوع بالضمة الظاهرة

C . مفعول به منصوب بالفتحة المقدرة على الألف

D . مفعول به منصوب بالفتحة الظاهرة

E . خبر مرفوع بالضمة المقدرة على الألف

١٠. التنوين في {مررت بسيبويهِ}

A . تنوين التنكير إذا قصد به شخص معين

B . تنوين التنكير إذا قصد به شخص غير معين

C . تنوين المقابلة إذا قصد به شخص معين

D . تنوين المقابلة إذا قصد به شخص غير معين

E . تنوين العوض إذا قصد به شخص معين

١١. كل ما يأتي لا يبين عن الاسم المقصور إلا:

A . هو الاسم الذي ينتهي بياء لازمة قبلها فتحة

B . هو الاسم الذي ينتهي بألف لازمة قبلها فتحة

C . هو الاسم الذي ينتهي بياء مقدرة قبلها كسرة

D . هو الاسم الذي ينتهي بألف مقدرة قبلها كسرة

E . هو الاسم الذي ينتهي بياء لازمة قبلها فتحة

١٢. إعراب كلمة "ليلى" في قولك {ذاكرت ليلَى الدرس}

A . فاعل مرفوع بالضمة الظاهرة

B . فاعل مرفوع بالضمة المقدرة لأنه من الأسماء المقصورة

C . مفعول به منصوب بالفتحة الظاهرة

D . فاعل مرفوع بالضمة المقدرة لأنه من الأسماء المنقوصة

E . فاعل منصوب بالفتحة المقدرة

١٣. كل ما يأتي أقسام المعارف إلا:

A . الضمائر

B . المحلى بالألف واللام

C . الاسم الموصول

D . الأسماء الستة

E . أسماء الإشارة

١٤. علامة جزم الفعل المعتل الآخر:

A . الكسرة

B . السكون

C . الكسرة المقدرة

D . حذف حرف العلة

E . حذف الكسرة

١٥. علامة الرفع في الكلمة التي تحتها خط في الآية الآتية هي:

{إن هذا أخي له تسع وتسعون نعجة}

A . النون . B . الضمة . C . الواو

D . التاء . E . الضمة المقدرة

١٦. إعراب "يذهبن" في قوله تعالى: {إن الحسنات يذهبن السيئات} الآية.

A . فعل مضارع مبني على حذف حرف العلة

B . فعل مضارع مبني على حذف النون

C . فعل مضارع مبني على الفتحة

D . فعل مضارع مبني على السكون

E . فعل مضارع مبني على الضم

١٧. كلمة "أخ" إذا ثنيت تكون:

A . أخان

B . أختان

C . أخيان

D . أخيان

E . أخوان

١٨. كلمة "الرامي" إذا جمعت تكون:

A . الرماة

B . الرامون

C . الراميون

D . الراموون

E . الرمية

١٩ . يبنى الفعل الماضي على الفتح إذا اتصلت به:

I . ألف الإثنين

II . واو الجماعة

III . تاء التأنيث الساكنة

IV . ياء المخاطبة

V . نون النسوة

A . I , II , III

B . I , IV , V

C . I , III , IV

D . I , III

E . I , IV

٢٠ . يبنى فعل الأمر على حذف النون إذا اتصلت به:

I . ألف الإثنين

II . نون النسوة

III . تاء التأنيث الساكنة

IV . واو الجماعة

V . ياء المخاطبة

A . I , II , III , IV

B . II , III , IV , V

III , II , I . C

IV , III , I . D

V , IV , I . E

٢١ . علامة نصب الأفعال الخمسة:

A . حذف حرف العلة

B . حذف النون

C . الياء

D . حذف حرف المضارع

E . الفتحة

٢٢ . فيما يأتي مواضع المسند إليه سوى:

A . فاعل الفعل التام وشبهه

B . نائب الفاعل

C . المبتدأ الذي له خبر

D . ما أصله مبتدأ

E . ما أصله خبر

٢٣ . ما وجه الشبه المحذوف في المثال الآتي:

{أنت كالبحر}

A . الضياء

B . الإشراق

C . الحسن

D . السماحة

E . الإحمرار

٢٤ . املأ الفراغ الآتي بالكلمة المناسبة:

{كلام فلان ك_____ في الحلاوة}

A . الملح

B . البن

C . الشهد

D . الشعر

E . الجبنة

القسم الثاني : المهارات اللغوية (التعبير والترجمة)

الطبيعة جميلة ___(٢٥)___ بآلاء الله وخيراته. عاش فيها الإنسان ينعم بما أفاء الله ___(٢٦)___ من نبت وزهر وثمر وطير وحيوان ، ومن هواء يتنسم فيه ومعه أنسام الحياة ومن ___(٢٧)___ وجداول وأنهار ___(٢٨)___ بمائها ___(٢٩)___ ويروي زرعه وماشيته ، وبحار يسعد بشواطئها ويغنى بثرواتها.

وكانت الطبيعة له مسرح إبداع ، و___(٣٠)___ جمال ، وميدان أرزاق، في كل شيء منها يد الله ، وفي كل خلق بها آية من آياته.

ولكن هذه الطبيعة بدأت ___(٣١)___ البشرية ، ويزيد شغلها حتى صار ___(٣٢)___ قلقا. لقد شق الإنسان الأرض ___(٣٣)___ تربتها وسبر أغوارها كي تبوح له بما ___(٣٤)___ في جوفها من معادن ومناجم و___(٣٥)___ ، وركب الأنهار وغاص في أعماق البحار بحثا عن ثرواتها الحيوانية ، و___(٣٦)___ الهواء لطائراته ومزقه بالصواريخ والأقمار الصناعية ومراكب الفضاء.

وحقق له ما اصطنع من ذلك خيرا وثراء وارتقاء ، ولكنه لم ___(٣٧)___ من كدر وشر ، ولم ___(٣٨)___ على الطبيعة فتنتها وسحرها. فقد ___(٣٩)___ هذه البيئة مما كانت تحظى به من جمال.

٢٥ . A . حفيلة	٢٦ . A . له	٢٧ . A . أعين	٢٨ . A . يروي
B . حافل	B . عليها	B . عينين	B . تشرب
C . حفلة	C . عليهم	C . عيون	C . يسقي
D . مليء	D . عليه	D . عين	D . يرتوي
E . حافلة	E . لهم	E . عينان	E . تسقي

٢٩ . A . العاذب	٣٠ . A . مجتلى	٣١ . A . تشتغل	٣٢ . A . تواجسا
B . العذب	B . مجالى	B . تشاغل	B . توجسا
C . الأعذب	C . اجتلى	C . تشغل	C . متوجسا
D . العذبة	D . جانب	D . يشغل	D . يتوجس
E . الطيبة	E . ناحية	E . يتشاغل	E . توجيسا

٣٣ . A . ينبت	٣٤ . A . سكن	٣٥ . A . كنائز	٣٦ . A . خضع
B . تزارع	B . يسكن	B . كتر	B . اذلل
C . تنبت	C . يخفى	C . كنوز	C . خضعت
D . يستنبت	D . استكن	D . كناوز	D . ذلل
E . يستخدم	E . ساكن	E . كترة	E . ذللت

٣٧ . A . يخلص	٣٨ . A . يحافظ	٣٩ . A . تلويث
B . يستخلص	B . ينظف	B . تلوث
C . يخاف	C . يحفظ	C . تلوثت
D . يحفظ	D . يدافع	D . كدر
E . يحافظ	E . يرعى	E . تعكر

٤٠. ترجم هذه العبارة إلى اللغة الملايوية:

(المصانع تجمع نفياتها حولها أو على مقربة منها)

A . Kilang-kilang perindustrian telah mengumpulkan sisa-sisa buangannya di sekelilingnya atau berdekatan dengannya.
B . Kilang-kilang industri mencampakkan sisa-sisa toksidnya di persekitarannya dan kawasan yang dekat dengannya.
C . Kilang-kilang perindustrian telah membuang sisa-sisa buangannya di sekitarnya dan di kawasan yang hampir dengannya.
D . Kilang-kilang perindustrian mengumpulkan bahan-bahan buangannya di persekitarannya atau berdekatan dengannya.
E . Kilang-kilang industri telah membuat kumpulan sisa-sisa toksidnya di sekelilingnya atau berdekatan dengannya.

٤١ . أحسن ترجمة للعبارة التالية :

(الإسلام دين الفطرة الإنسانية بلا شك)

A . Agama Islam adalah agama Fithrah untuk manusia tanpa syak wasangka.
B . Agama Islam merupakan agama Fithrah manusia semenjak azali lagi.
C . Agama Islam merupakan agama Fithrah manusia tanpa ragu-ragu lagi.
D . Tiada syak lagi bahawa agama Islam adalah merupakan agama Fithrah manusia.
E . Tanpa ragu-ragu Agama Islam adalah agama Fithrah manusia

٤٢ . أحسن ترجمة للعبارة التالية هي:

(الإسلام يحافظ على شخصية الفتاة في الزواج)

A . Agama Islam menjaga setiap individu wanita dalam soal nikah kahwin.
B . Agama Islam amat menitikberatkan soal kehidupan wanita dalam perkahwinan.
C . Agama Islam memelihara keperibadian wanita dalam urusan perkahwinan.
D . Islam adalah sebuah agama yang menitikberatkan soal perkahwinan wanita.
E . Agama Islam amat memelihara watak pemudi dalam soal nikah kahwin.

٤٣ . أحسن ترجمة للعبارة الآتية هي:

(هل مستقبل الإنسانية سيكون مظلما في المستقبل)

A . Adakah masa depan kemanusian menjadi gelap gulita.
B . Adakah masa depan manusia akan menjadi gelap.
C . Adakah kemanusian akan menjadi gelap di masa hadapan.
D . Adakah masa depan kemanusian akan menjadi gelap di masa hadapan.
E . Adakah kegelapan manusia akan terjadi di masa hadapan.

٤٤ . أحسن ترجمة للعبارة التالية هي:

(عالم أطفالنا سيختلف في المستقبل عن عالمنا اليوم)

A . Alam kanak-kanak di masa hadapan berbeza dengan alam kita pada hari ini.
B . Ada perbezaan antara alam kanak-kanak dan alam kita di masa hadapan.
C . Akan berbeza antara alam kanak-kanak dan alam kita pada hari ini dan masa hadapan.
D . Akan wujud perbezaan yang ketara antara alam kanak-kanak dan alam kita pada hari ini.
E . Alam kanak-kanak di masa hadapan akan berbeza dengan alam kita hari ini.

٤٥ . ترجم العبارة التالية إلى اللغة العربية:

(وانيتا مروفاكن سبهاكين درفد مشاركة مانسي)

A . النساء فرع من فروع المجتمع الإنساني.

B . المرأة نصف المجتمع الإنساني

C . المرأة جزء من أجزاء مجتمع الإنسان

D . المرأة جزء مجتمع الإنسان

E . النساء هن من المجتمع الإنساني.

٤٦ . أحسن ترجمة للعبارة الآتية هي:

(فارا فلاوت دمينتا مليهارا أداب٢ ملاوت)

A . الزائرون عليهم أن يتبعوا الآداب للزيارة

B . أداب الزيارة لا بد للزائرين اتباعها

C . يرجى من جميع الزائرين أن يراعوا آداب الزيارة

D . للزائرين آداب خاصة لهم للزيارة

E . على الزائرين مراعات الآداب الزيارة

EX1201 (sambungan)

٤٧ . أحسن ترجمة للعبارة الآتية هي:

(هوجن يع لبت تله مععاقبتكن كسسقكن لالو لينتس)

A . ازدحام الطرق قد سبب المطر الغزير

B . قد سبب المطر الغزير ازدحام في الطريق

C . يزدحم الطرق وذلك بعد المطر الغزير

D . ذلك المطر قد جاء بعد ازدحام الطرق

E . المطر الغزير سبب الازدحام الطرق

٤٨ . أحسن ترجمة للعبارة الآتية هي:

(علمو فعتاهوان منوجو كأره كماجوان يع برتروسن)

A . العلم في تقدم دائما

B . العلوم في تقدم دائم

C . العلم يسير إلى التقدم الدائم

D . العلوم في رفاهية مستمرة

E . العلم في تقدم مستمر

٤٩ . هات مرادف كلمة "آلاء":

A . علامات

B . عناوين

C . نعم

D . براهين

E . دلائل

٥٠. هات مرادف كلمة "استكن"

A . اختفى

B . اختبأ

C . اصطفى

D . تكشف

E . تغلغل

القسم الثالث: المقال

اكتب مقالا تتحدث فيه عن أحد الموضوعات الآتية بحيث لا يقل عدد كلماته عن مائتى كلمة:

١ . دور الشباب في بناء المستقبل المشرق للدين والوطن.

٢ . النظر في الكون سبيل الإيمان.

٣ . أهمية التعاون والإتحاد للأمة الإسلامية.

٤. قمت برحلة إلى منطقة جبلية ، صف عن روعة المناظر وجمال الشلالات فيها.

٥. جمال الحياة في القرى.

(١٥ درجة)

****

A.3 Questionnaires

A.3.1 Questionnaire for teachers

PENILAIAN OLEH GURU BAHASA ARAB TERHADAP ITEM UJIAN

BULATKAN nombor yang sesuai menurut pandangan anda berdasarkan skala berikut:

skala: 1 = sangat sesuai/sangat berkaitan/sangat jelas
2 = sesuai/berkaitan/jelas
3 = kurang sesuai/kurang berkaitan/kurang jelas
4 = amat tidak sesuai/tidak berkaitan langsung/sangat kabur

UJIAN BACAAN DAN KEFAHAMAN

A. KULIT SOALAN

Arahan di kulit soalan	1	2	3	4

B. TEKS UJIAN

(a) Bahagian Pertama (Aneka pilihan)

i. Teks 1

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

ii. Teks II

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

iii. Teks III dan IV

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

(b) Bahagian Dua (True-false)

i. Teks I

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

ii. Teks II

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

iii. Teks III

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

iv. Teks IV

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

iv. Teks V

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

(c) Bahagian Tiga (Cloze test)

Teks ujian:

a. Perbendaharaan kata	1	2	3	4
b. Struktur ayat	1	2	3	4
c. Kesesuaiannya dengan keupayaan calon	1	2	3	4
d. kandungan isi (content)	1	2	3	4
e. Panjang teks	1	2	3	4
f. Ketidakserasian budaya (cultural bias)	1	2	3	4

C. SOALAN-SOALAN

(a) Bahagian Satu (aneka pilihan)

Soalan 1-10

a. Arahan yang diberikan	1	2	3	3
b. Kejelasan soalan	1	2	3	4
c. Ketepatan soalan dengan teks	1	2	3	4
d. Aras soalan (level)	1	2	3	4
e. Format soalan	1	2	3	4
f. Keserasian calon dengan bentuk soalan	1	2	3	4
g. Masa yang diperuntukkan	1	2	3	4

(b) Bahagian Dua (True-false)

Soalan 11-30

a. Arahan yang diberikan	1	2	3	3
b. Kejelasan soalan	1	2	3	4
c. Ketepatan soalan dengan teks	1	2	3	4
d. Aras soalan (level)	1	2	3	4
e. Format soalan	1	2	3	4
f. Keserasian calon dengan bentuk soalan	1	2	3	4
g. Masa yang diperuntukkan	1	2	3	4

(c) Bahagian Tiga (Cloze)

Soalan 31-75

a. Arahan yang diberikan	1	2	3	3
b. Kejelasan soalan	1	2	3	4
c. Ketepatan soalan dengan teks	1	2	3	4
d. Aras soalan (level)	1	2	3	4
e. Format soalan	1	2	3	4
f. Keserasian calon dengan bentuk soalan	1	2	3	4
g. Masa yang diperuntukkan	1	2	3	4

PENILAIAN OLEH GURU BAHASA ARAB TERHADAP ITEM UJIAN

BULATKAN nombor yang sesuai menurut pandangan anda berdasarkan skala berikut:

skala: 1 = sangat sesuai/sangat berkaitan/sangat jelas
2 = sesuai/berkaitan/jelas
3 = kurang sesuai/kurang berkaitan/kurang jelas
4 = amat tidak sesuai/tidak berkaitan langsung/sangat kabur

UJIAN NAHU BAHASA ARAB

A. KULIT SOALAN

Arahan di kulit soalan	1	2	3	4

B. SOALAN-SOALAN

(a) Bahagian Satu (Aneka pilihan)

Soalan 1—45:

a. Arahan yang diberikan	1	2	3	3
b. Kejelasan soalan	1	2	3	4
c. Ketepatan soalan dengan teks	1	2	3	4
d. Aras soalan (level)	1	2	3	4
e. Format soalan	1	2	3	4
f. Keserasian calon dengan bentuk soalan	1	2	3	4
g. Masa yang diperuntukkan	1	2	3	4
h. Pemilihan ayat-ayat soalan	1	2	3	4

(a) Bahagian Dua (True-false)

Soalan 46—50:

a. Arahan yang diberikan	1	2	3	3
b. Kejelasan soalan	1	2	3	4
c. Ketepatan soalan dengan teks	1	2	3	4
d. Aras soalan (level)	1	2	3	4
e. Format soalan	1	2	3	4
f. Keserasian calon dengan bentuk soalan	1	2	3	4
g. Masa yang diperuntukkan	1	2	3	4
h. Pemilihan ayat-ayat soalan	1	2	3	4

Komen Tambahan:

PENILAIAN OLEH GURU BAHASA ARAB TERHADAP ITEM UJIAN

BULATKAN nombor yang sesuai menurut pandangan anda berdasarkan skala berikut:

skala: 1 = sangat sesuai/sangat berkaitan/sangat jelas
2 = sesuai/berkaitan/jelas
3 = kurang sesuai/kurang berkaitan/kurang jelas
4 = amat tidak sesuai/tidak berkaitan langsung/sangat kabur

UJIAN MENULIS DALAM BAHASA ARAB

A. KULIT SOALAN

Arahan di kulit soalan	1	2	3	4

B. SOALAN

a. Kejelasan soalan	1	2	3	4
b. Ketepatan soalan dengan sukatan	1	2	3	4
c. Aras soalan (level)	1	2	3	4
d. Format soalan	1	2	3	4
e. Keserasian calon dengan bentuk soalan	1	2	3	4
f. Masa yang diperuntukkan	1	2	3	4
g. Kesesuaian tajuk esei dengan calon	1	2	3	4
h. Minat calon terhadap tajuk esei	1	2	3	4
i. Pemilihan isi-isi penting	1	2	3	4

PENILAIAN OLEH GURU BAHASA ARAB TERHADAP ITEM UJIAN

BULATKAN nombor yang sesuai menurut pandangan anda berdasarkan skala berikut:

skala: 1 = sangat sesuai/sangat berkaitan/sangat jelas
2 = sesuai/berkaitan/jelas
3 = kurang sesuai/kurang berkaitan/kurang jelas
4 = amat tidak sesuai/tidak berkaitan langsung/sangat kabur

UJIAN KEMAHIRAN EJAAN DALAM BAHASA ARAB

A. KULIT SOALAN

Arahan di kulit soalan	1	2	3	4

B. SOALAN

a. Kejelasan teks yang dirakam	1	2	3	4
b. Ketepatan soalan dengan sukatan	1	2	3	4
c. Aras soalan (level)	1	2	3	4
d. Format soalan	1	2	3	4
e. Keserasian calon dengan bentuk soalan	1	2	3	4
f. Tempoh pemberhentian seketika (pause)	1	2	3	4
g. Pemilihan ayat-ayat soalan	1	2	3	4
h. Panjang teks ujian	1	2	3	4
i. Kesesuaian kandungan dengan calon	1	2	3	4

A.3.2 Questionnaire for students

PENIALAIAN KENDIRI TERHADAP PENGUASAAN BAHASA ARAB

Nama: _____ Fakulti: _____

Gred diperolehi untuk kertas Bahasa Arab STPM/Pra Akademi

Kertas 1: _____ Kertas 2: _____ Kertas 3: _____

Gred keseluruhan/Kertas 4: _____

Bulatkan nombor yang bersetentang dengan penyataan yang diberikan untuk jawapan anda kepada soal-selidik berikut mengunakan skala di bawah:

1 = sangat lemah
2 = lemah
3 = baik
4 = sangat baik

i. Kemahiran bacaan dan pemahaman

Nyatakan penilaian anda terhadap diri anda sendiri untuk kemahiran bacaan dan pemahaman menggunakan skala yang telah di berikan di atas:

1. memahami teks yang dibaca	1	2	3	4
2. menentukan baris perkataan di dalam teks yang di baca	1	2	3	4
3. memahami makna perkataan dalam teks yang di baca	1	2	3	4
4. merumuskan isi penting penulis teks tersebut	1	2	3	4
5. memahami teks yang tidak berkaitan bidang pengajian anda (spt. Sastera, Sains, Fiksyen dsb.nya)	1	2	3	4
6. memahami teks yang berkaitan bidang pengajian anda (spt. Syariah, Usuluddin, dsb.nya)	1	2	3	4
7. menjawab soalan berbentuk aneka pilihan	1	2	3	4
8. menjawab soalan berbentuk betul-salah (true-false)	1	2	3	4
9. menjawab soalan berbentuk 'cloze test'	1	2	3	4

ii. Kemahiran Naḥu dan Ṣarf

Nyatakan penilaian anda terhadap diri anda sendiri untuk kemahiran Nahu dan Sarf menggunakan skala yang telah di berikan di atas:

1. memahami tajuk-tajuk nahu berikut:				
a. *mabni wa mu`rab*	1	2	3	4
b. *nakirah wa ma`rifa*	1	2	3	4
c. *mubtada' wa khabar*	1	2	3	4
d. *kana wa akhawatuha*	1	2	3	4
e. *inna wa akhawatuha*	1	2	3	4
2. memahami tajuk-tajuk Sarf berikut:				
a. *mufrad, muthanna, jam`*	1	2	3	4
b. *jāmid wa mutaṣarrif*	1	2	3	4
c. *jāmid wa mushtaq*	1	2	3	4
3. menjawab soalan berbentuk aneka pilihan	1	2	3	4
4. menjawab soalan berbentuk betul salah	1	2	3	4
5. menulis jawapan yang betul terhadap jawapan salah	1	2	3	4

iii. Kemahiran ejaan

Nyatakan penilaian anda terhadap diri anda sendiri untuk kemahiran ejaan menggunakan skala yang telah di berikan di atas:

1. menulis sepenuhnya apa yang didengari	1	2	3	4
2. menentukan perkataan yang didengari itu terdiri dari satu perkataan atau lebih	1	2	3	4

3. menulis perkataan yang bersambung dengan alif lam qamariyya	1	2	3	4
4. menulis perkataan yang bersambung dengan alif lam shamsiyya	1	2	3	4
5. membezakan antara huruf ( ) dan ( )	1	2	3	4
6. membezakan antara huruf ( ) dan ( )	1	2	3	4
7. membezakan antara huruf ( ) dan ( )	1	2	3	4
8. menentukan samada perkataan yang didengari mempunyai harakat panjang atau pendek	1	2	3	4

iv. Kemahiran menulis karangan

Nyatakan penilaian anda terhadap diri anda sendiri untuk kemahiran menulis karangan menggunakan skala yang telah di berikan di atas:

1. menyusun fakta isi karangan terhadap tajuk yang diberikan	1	2	3	4
2. menulis dengan jelas apa yang hendak diperkatakan	1	2	3	4
3. tidak melakukan kesalahan nahu dalam penulisan	1	2	3	4
4. menguasai perbendaharaan kata yang baik untuk isi karangan yang ditulis	1	2	3	4
5. menguasai ejaan dengan baik bagi setiap perkataan yang hendak ditulis	1	2	3	4

Terima kasih di atas kerjasama anda

Appendix B Data

B.1 Item Facility for the Reading, Grammar, and Dictation tests at the AIS (N=413)

Item Facility (IF) for the Reading test for samples from AIS (N=413)

Item no.	Item Facility (IF)
1	.70
2	.41
3	.26
4	.39
5	.69
6	.72
7	.87
8	.71
9	.75
10	.93
11	.81
12	.90
13	.73
14	.62
15	.61
16	.61
17	.66
18	.44
19	.41
20	.94
21	.24
22	.57
23	.60
24	.68
25	.30
26	.90
27	.42
28	.39
29	.78
30	.55
31	.08
32	.10
33	.13
34	.80
35	.57
36	.41
37	.38
38	.15
39	.26
40	.18
41	.05
42	.41
43	.03
44	.25
45	.63
46	.60
47	.68
48	.15
49	.05
50	.06
51	.05
52	.05
53	.20
54	.14
55	.23
56	.37
57	.19
58	.29
59	.46
60	.57
61	.09
62	.41
63	.59
64	.31
65	.69
66	.02
67	.03
68	.05
69	.58
70	.20
71	.14
72	.12
73	.40
74	.23
75	.04

Item Facility for the Grammar test for samples from AIS (N=413)

Item no.	Item Facility (IF)
1	.62
2	.68
3	.74
4	.26
5	.32
6	.69
7	.60
8	.47
9	.44
10	.55
11	.59
12	.56
13	.16
14	.35
15	.64
16	.56
17	.70
18	.24
19	.20
20	.57
21	.46
22	.81
23	.55
24	.36
25	.78
26	.19
27	.60
28	.30
29	.50
30	.26
31	.26
32	.58
33	.32
34	.26
35	.64
36	.44
37	.54
38	.28
39	.54
40	.36
41	.16
42	.24
43	.52
44	.44
45	.49
46	.72
47	.65
48	.84
49	.79
50	.54

Item Facility for the Dictation test for samples from AIS (N=413)

Item no.	Item Facility (IF)
1	.63
2	.68
3	.80
4	.54
5	.84
6	.65
7	.81
8	.53
9	.50
10	.50
11	.44
12	.24
13	.81
14	.32
15	.55
16	.33
17	.68
18	.43
19	.86
20	.69
21	.37
22	.23
23	.68
24	.12
25	.50

B.2 Item Discrimination for the Reading, Grammar, and Dictation test at the AIS

Item discrimination (ID) for the Reading test for samples from AIS

Item	IF (upper)	IF (lower)	ID
1	.90	.36	.54
2	.81	.09	.72
3	.42	.17	.25
4	.56	.19	.37
5	.93	.35	.58
6	.94	.43	.51
7	.97	.78	.19
8	.98	.39	.59
9	.86	.58	.28
10	1.00	.81	.19
11	.91	.64	.27
12	.95	.82	.13
13	.94	.62	.33
14	.89	.44	.45
15	.77	.46	.31
16	.95	.36	.59
17	.91	.35	.56
18	.66	.25	.41
19	.54	.37	.17
20	1.00	.87	.13
21	.38	.16	.22
22	.68	.58	.10
23	.79	.43	.36
24	.88	.53	.35
25	.39	.26	.13
26	.97	.72	.25
27	.58	.33	.25
28	.52	.33	.19
29	.76	.69	.07
30	.89	.20	.69
31	.28	.01	.27
32	.25	.04	.21
33	.29	.01	.28
34	.96	.61	.35
35	.88	.24	.64
36	.76	.15	.61
37	.67	.13	.54
38	.35	.02	.33
39	.68	.03	.65
40	.36	.07	.29
41	.12	.01	.11
42	.67	.19	.48
43	.05	.01	.04
44	.67	.04	.63
45	.93	.33	.60
46	.87	.27	.60
47	.93	.35	.58
48	.49	.01	.48
49	.11	.02	.10
50	.18	.00	.18
51	.16	.01	.15
52	.15	.02	.13
53	.35	.11	.24
54	.38	.02	.36
55	.39	.10	.29
56	.64	.17	.47
57	.32	.05	.27
58	.74	.04	.70
59	.97	.09	.88
60	.95	.13	.82
61	.25	.02	.23
62	.85	.10	.75
63	.96	.16	.80
64	.59	.13	.46
65	.83	.51	.32
66	.04	.01	.03
67	.02	.00	.02
68	.10	.01	.09
69	.89	.30	.59
70	.39	.02	.37
71	.38	.02	.36
72	.23	.01	.22
73	.77	.15	.62
74	.53	.03	.50
75	.15	.01	.14

Item discrimination (ID) for the Grammar test for samples from AIS (N=413)

Item	IF (upper)	IF (lower)	ID
1	.85	.35	.50
2	.94	.39	.55
3	.93	.51	.42
4	.51	.09	.42
5	.69	.08	.61
6	.91	.46	.45
7	.83	.30	.53
8	.66	.28	.38
9	.74	.22	.52
10	.87	.23	.64
11	.90	.36	.54
12	.75	.29	.46
13	.12	.24	-.12
14	.48	.26	.22
15	.99	.23	.76
16	.95	.10	.85
17	.92	.45	.47
18	.56	.14	.42
19	.48	.06	.42
20	.87	.26	.61
21	.70	.29	.41
22	.98	.56	.42
23	.76	.35	.41
24	.76	.08	.68
25	.98	.58	.40
26	.32	.13	.19
27	.82	.35	.47
28	.50	.18	.32
29	.76	.30	.46
30	.51	.12	.39
31	.47	.12	.35
32	.92	.26	.66
33	.49	.17	.32
34	.48	.14	.34
35	.90	.35	.55
36	.61	.30	.31
37	.91	.31	.60
38	.54	.11	.43
39	.78	.38	.40
40	.73	.17	.56
41	.39	.07	.32
42	.46	.14	.32
43	.82	.25	.57
44	.83	.15	.68
45	.90	.25	.65
46	.94	.46	.48
47	.93	.27	.66
48	.97	.67	.30
49	.93	.66	.27
50	.89	.25	.64

Item discrimination (ID) for the Dictation test for samples from AIS

Item	IF (upper)	IF (lower)	ID
1	.90	.27	.63
2	.88	.50	.38
3	.98	.54	.44
4	.97	.11	.86
5	.99	.50	.49
6	.96	.28	.68
7	.99	.54	.45
8	.94	.15	.79
9	.95	.10	.85
10	.91	.12	.79
11	.90	.13	.77
12	.56	.01	.55
13	1.00	.48	.52
14	.73	.05	.68
15	.90	.13	.77
16	.59	.17	.42
17	.98	.31	.67
18	.68	.21	.47
19	.97	.74	.23
20	.99	.27	.72
21	.87	.04	.83
22	.38	.08	.30
23	.83	.50	.33
24	.33	.00	.33
25	.83	.14	.69

B.3 Descriptive statistics for the placement test for session 1999/00

Descriptive statistics for the placement test for session 1999/00 (N=555)

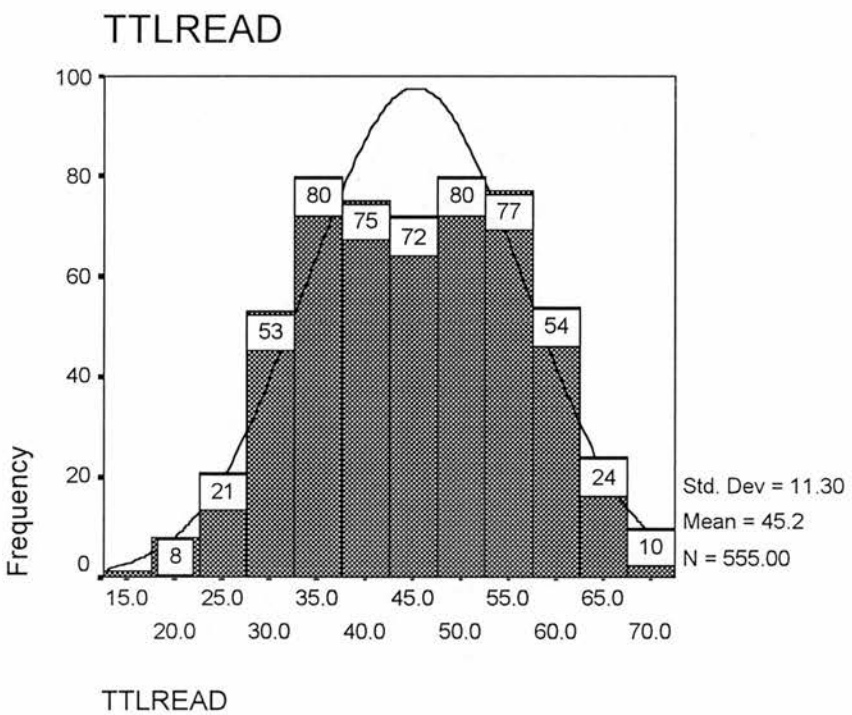
(a) The Reading test

Statistics

TTLREAD

N	Valid	555
	Missing	0
Mean		45.19
Median		46.00
Mode		35[a]
Std. Deviation		11.30
Variance		127.58
Range		54
Minimum		17
Maximum		71
Sum		25081

a. Multiple modes exist. The smallest value is shown



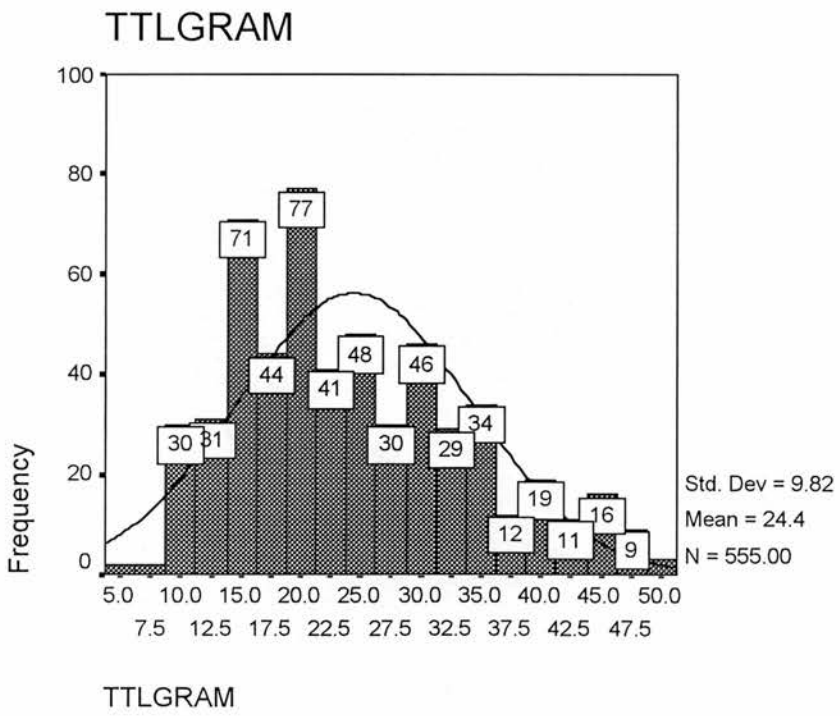
(b) The Grammar test

Statistics

TTLGRAM

N	Valid	555
	Missing	0
Mean		24.40
Median		23.00
Mode		16[a]
Std. Deviation		9.82
Variance		96.52
Range		45
Minimum		6
Maximum		51
Sum		13540

a. Multiple modes exist. The smallest value is shown



Appendix C Recorded voice for the Dictation Test (in tape)